Occupational Variety and Economic Development

Anton Heil, Gabriel Leite-Mariante

January 7, 2026

Click for latest version

Abstract

The paper introduces, formalizes, and tests the idea of occupation-based linkages between industries. Industries are linked through their common requirement of specialised skills. Occupations capture those specific skills and allow us to measure them. We first validate this intuition by documenting that occupational variety systematically rises with economic development (and industrial variety) across countries and across regions within Brazil. We then use detailed micro-data from Brazil to show that new industries and occupations emerge jointly, forming a bi-directional network where industries hire multiple, overlapping occupations and occupations work in multiple, overlapping industries. To explain these facts, we construct a model of industrialisation with occupation-based linkages. Contrary to standard models where industries would compete over specialised workers, we show that positive externalities between industries can arise, e.g. if there are matching frictions and entry of one industry thickens the labour market for the occupations it requires. Finally, using a shift-share instrument approach, we show that occupation-based linkages are positive and of a similar magnitude as traditional input-output linkages. In line with the model, a region's position in the industry-occupation network predicts the direction of industrialisation and regional growth. The results imply that education/training and industrial policies are complements and that targeting bottleneck occupations can unlock cascades of diversification.

1 Introduction

The diversification of economic activities is one of the key features of economic development. A well known example of this is that richer economies have a larger variety of industries. This can be seen in the splintering of the workforce across sectors (Imbs and Wacziarg, 2003) or in the increasing variety and complexity of export products (Hidalgo et al., 2007). Another striking form of diversity is the variety of available occupations. Workers in the poorest places typically choose between a handful of jobs while rich economies offer a large variety of specialised occupations. This pattern holds across countries, over time and across regions within the same country (see Figure 1).[1]

The path and speed of industrialisation - the emergence of new economic activities - has traditionally been explained by invoking externalities between industries. For example, classical theories of development have highlighted linkages between industries through aggregate demand (Rosenstein-Rodan, 1943) or input-output networks (Hirschman, 1958).

This paper provides a novel approach to explaining the expansions of economic activities, which links the rise in industrial to occupational diversification. Specifically, we introduce a new type of linkage between industries whereby two industries are linked through the fact that they both require skilled workers from the same occupation. For example, the food processing and the pharmaceutical industry might not have much in common, except for the fact that they both hire chemists. Yet, this common occupational requirement means that entry and growth of one industry affects the other industry through its effect on the labour market for chemists. We refer to this as an *occupation-based linkage*. The paper formalises this concept in a model of industrialisation with occupations and test it using granular labour market records from Brazil over a period of 19 years. In doing so, we provide a novel, labour-centred perspective on industrialisation, with far reaching implications for industrial, training, and education policy.

The paper is structured in three parts. The first is to document the robust relationship between occupational variety and economic activity (section 3). For this, we assemble a novel dataset on occupational variety from national census micro-data and historical censuses. We use the census micro data to decompose the aggregate relationship and analyse the occupational structure within and between different subgroups of the population. These decompositions show that the rise in occupational variety can be partly accounted for by compositional shifts driven by other, well-known dimensions of structural transformation, such as an increase in formal education, a decline in agriculture shares, urbanisation, and a shift from self-employment to wage work. However, none of these transformations fully accounts for the increase in occupational variety, suggesting that it might be an important dimension of structural change in its own right. We then use administrative micro-data on the universe of labour contracts in Brazil between 2003-2021 to study the relationship between industrial and occupational variety. Brazil serves as a

[1] For details see the discussion of this Figure in section 3 below.

useful case study since due to its large regional inequality, its regions span the economic structure of most countries of the world. We show that across Brazilian regions, new industries and new occupations emerge jointly, suggesting that entry of new industries - rather than within industry diversification - is the main force behind the increase in occupational variety. We also show that industries typically hire a variety of workers from different occupations, and that occupational requirements often overlap across industries. On the other hand, workers from the same occupation often work in many different industries, and the industry employment profiles of occupations often overlap. This means that we can picture the relationship between industries and occupations as a bi-directional network, where industries are linked through their use of common occupations and occupations are linked through their employment in common industries.

To interpret these facts, the second part of the paper presents a model of industrialisation with specialised occupations and heterogeneous workers (section 4). Firms require fixed bundles of specialised labour in different proportions, depending on their industry. They only enter the market if the required labour is available locally at a low enough wage. This means that entry of new industries can be constrained by a lack of specifically skilled workers. As skilled occupations become more available in the local labour market, industries using these occupations are more likely to enter. Workers, on the other hand, have to undergo costly training in order obtain the skills required for an occupation. They weigh the cost of training in an occupation against the wage and the probability of being hired. Hiring is subject to matching frictions: if only few active firms can hire an occupation, the matching probability is low.

This set-up can generate a coordination failure where firms don't enter because the specialised labour they require is not available (or prohibitively expensive) and workers don't train in specialised occupations because the probability to be hired is too low. New industries only emerge if this coordination problem can be overcome. This can happen through occupational linkages. Wages and hiring probability are complements in the worker's training decision. As active firms raise hiring probabilities in the occupations they use, they make these occupations more attractive to potential trainees, thus lowering effective hiring costs for other firms considering entry in the future. This illustrates how an occupation-based linkage can generate a positive externality: if two industries share occupations, entry of one facilitates entry of the other. Entry thus propagates along the occupation–industry network.

At the industry level, this mechanism implies that industries which require few, low-skill, and already-active occupations tend to enter earlier, whereas industries that rely on many non-overlapping or particularly skill-intensive occupations arrive later, if at all. At the economy level, catch-up growth is driven by this extensive entry margin and is path dependent: locations sitting in denser parts of the overlap network diversify faster. Policy can accelerate the process by lowering training costs in bottleneck occupations, improving matching, or subsidizing strategically positioned industries that unlock multiple followers.

Finally, the model highlights the potential aggregate productivity gains that arise from occupational variety. Workers in our framework are horizontally diversified - they have a different potential talent for each possible occupation. When only a few occupations are available locally, much of this talent is lost - a form of misallocation that arises not because workers match to the wrong occupation, but because the optimal occupation for many of them does not yet exist. As the set of active occupations expands, workers can sort into tasks that fit their comparative advantage, thus unleashing previously wasted talent. Occupational variety thus expands the scope for allocative efficiency. Variety also strengthens incentives to acquire skills. Entering an occupation requires costly training, and workers are more willing to invest when they expect a job that rewards their particular strengths. A broader menu of occupations increases the chance that such a match is available, thus creating incentives for specialised training and shifting people out of subsistence into higher productivity employment. Through these two channels, the expansion of occupations increases effective labour and aggregate productivity at the extensive margin.

In the final part of the paper, we test for occupation-based linkages using the labour market data from Brazil (section 5). To do so, we construct a model-based measure of proximity between industries based on their use of common occupations in a benchmark region. This measure of occupational linkage predicts entry of new industries, suggesting a positive externality: entry is *more* likely when an industry's occupational requirement overlaps with the occupational composition of already active industries in the region. This effect is quantitatively important. A one standard-deviation increase in the occupational overlap measure increases the entry probability of a new industry by 15%. This effect is unchanged when controlling for local demand spillovers, industry-level technology shocks. It is robust to controlling for input-output linkages between entering and pre-existing industries, and of a similar magnitude to these traditional linkages. Further, we document a positive linkage effect across broad industry groups, alleviating concerns of correlated demand shocks.

To corroborate a causal interpretation of the occupation-based linkage effect, we construct a shift-share instrumental variable based on initial industry-region employment shares and exogenous, time-variant shifts at the industry level, such as aggregate exports of the goods produced by an industry. These results lend further credibility to the interpretation of a causal linkage effect through the labour market for common occupations.

Finally, aggregating the occupation-based industry entry potential at the regional level, we show that this measure predicts regional growth conditional on initial levels of economic activity. Overall, the findings are consistent with the idea that growth is path dependent and faster in more connected areas of the industry-occupation network, thus validating the model's comparative statics and highlighting the role of occupation-based linkages for industrialisation.

The presence of occupational linkages has several implications for industrial policy. The first is that a lack of specialised labour can undermine efforts to advance industries that rely on such workers. Training and education policy is therefore complementary to traditional industrial policy in the form of subsidised loans or tax breaks. Furthermore, addressing bottlenecks in the supply of widely-used occupations, for example through training subsidies or improved matching, can facilitate entry of new industries. Hirschman's (1953) original idea of targeting industries with strong input-output linkages, translates in our model to the recommendation of targeting industries with strong human capital spill-overs. These industries will typically be centrally located in the industry-occupation network. Support to those industries can encourage specialised training in many occupations that are in turn useful to many other industries, and thus stimulate further industrialisation. Finally, growth diagnostics (e.g. Hausmann et al. (2008)) should take the specific complementarities displayed in the industry-occupation network into account - low returns to education might occur when workers with specific training face low employment probabilities, and low returns to capital injections might occur when firms cannot find workers specialised in the particular skills they require.

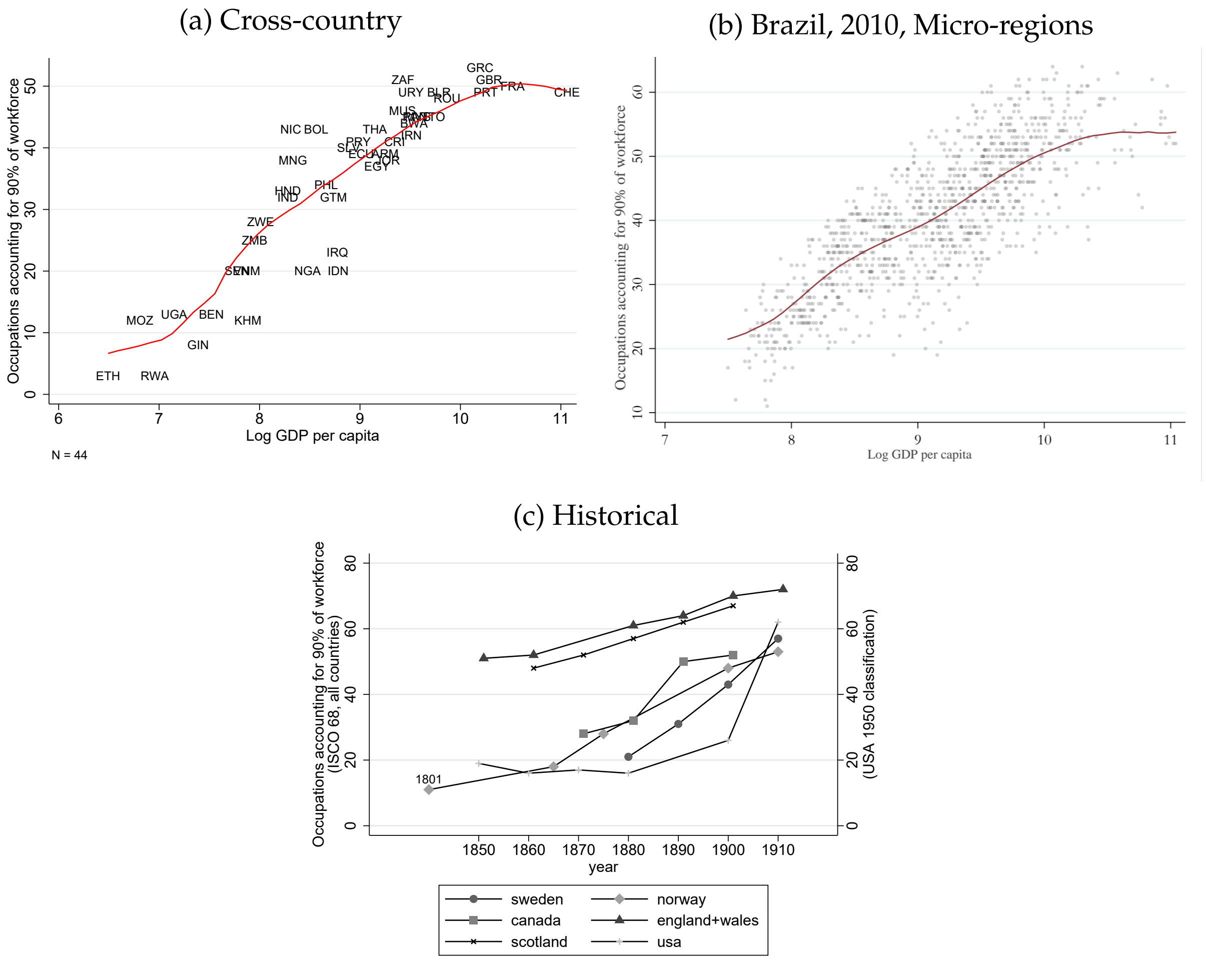
The paper links to the literature on industrialisation and industrial policy (Rosenstein-Rodan, 1943, Hirschman, 1958, Murphy et al., 1989, Matsuyama, 1991, Liu, 2019, Lane, 2025). We contribute to this literature by introducing an central role of specialised labour as a complementary input to industrialisation and highlighting ways to incorporate this insight into industrial policy. This literature has traditionally been divided into those who, following Rosenstein-Rodan (1943), emphasize balanced growth and a big-push approach to industrialisation, and those who, following Hirschman (1958) emphasize unbalanced growth along strategic sectors based in linkages. Our theory bridges these two traditions: first movers into new industries are crucial and can trigger further industrialisation in linked industries, but industries are linked through their common use of specialised labour. The coordination problem, highlighted in the big-push literature, is retained but relocated to the level of individual industries. The idea that different capabilities are required to produce different products, and that an economy can more easily move into new products that require similar capabilities has been previously advanced in the literature on economic complexity and the product space (Hidalgo et al., 2007, Hausmann et al., 2007, Hidalgo and Hausmann, 2009, Neffke and Henning, 2013). We contribute to this literature by providing a measurable micro-foundation of capabilities in the form of specialised labour in occupations. We also contribute to the empirics of that literature, by constructing a product space based on occupational linkages and showing that it explains observed patterns of industrialisation.

Our theory relies on previous work on complementarities in production (Kremer, 1993) and barriers to specialised training (Acemoglu, 1997). Particularly related are previous papers that show how low-development traps can arise from a low variety of intermediary inputs, which can be re-interpreted as types of specialised labour, or occupations (Rodriguez-Clare, 1996, Rodrik, 1996, Ciccone and Matsuyama, 1996). The labour supply side of our model follows previous work on occupational choice and comparative advantage - in particular Hsieh et al. (2019).

More broadly, the paper contributes to an emerging literature that uses micro-data to document and analyse structural transformations (Bandiera et al., 2022, Gollin and Kaboski, 2023) and to previous empirical work that has highlighted the importance of occupational variety for specialisation. For example, Papageorgiou (2022) documents that workers in larger cities have more occupational choice options and uses a structural model to show that this accounts for a third of the observed wage premium and greater inequality in larger cities. Tian (2021) uses the same Brazilian labour market data to show that firms in larger cities use more different occupations with resulting productivity gains. The idea that differentiated, specialised skills constitute an important form of human capital (beyond years of formal education) is also prominent in Jones (2008). We focus on a specific notion of specialisation: an expansion of the set of tasks performed in the economy because of the production of new goods and services, which allows workers to chose those tasks at which they have a comparative advantage. This differs from a notion of specialisation as a finer division of a given set of tasks (as e.g. in Chaney and Ossa (2013)).

2 Data

We use three sources of data on occupational variety. For cross-country and historical analysis we use census data that has been harmonised in the Integrated Public Use Microdata Series (IPUMS, 2020). For in-depth analysis of occupational patterns and regional development, we use Brazilian employer-employee matched data from the Relação Anual


(a) Cross-country

(b) Brazil, 2010, Micro-regions

(c) Historical

Figure 1: Economic growth and occupational variety

de Informaçoes (RAIS).

Integrated Public Use Microdata Series (IPUMS)

IPUMS is a a large data collection and harmonization exercise conducted by the University of Minnesota's Institute for Social Research and Innovation (IPUMS (2020)).[2] It contains household-level microdata from National Censuses covering over 100 countries. The data typically are a 1-%10% random sample of the population census.[3]

For 44 countries the census microdata contains detailed information on individual's work such as labour force participation, employment status, and an occupation code. Occupations are encoded based on the classification used by the national statistical agency. Where possible, a harmonized occupation variable has been created using the ILO's International Classification of Occupations (ISCO). In most cases the 1988 version of ISCO is used.[4] The ISCO 1988 classification contains 9 major groups (plus armed forces) and 116 possible minor groups at the 3-digit level.[5]

Historical census micro-data

We also retrieve historical census data from IPUMS to look at the evolution of now-industrialised countries. These cover selected countries that conducted censuses in the late 19th and early 20th century (Sweden, Norway, Canada, England and Wales, Scotland, and the USA). Occupations are recode into the ISCO 1968 version for all countries except the USA which uses a US-specific 1950 classification.

Relação Anual de Informaçoes Sociais (RAIS)

RAIS is a comprehensive labour market database maintained by Brazil's Ministério do Trabalho e Emprego (MTE) (Ministry of Labor and Employment). It contains the universe of formal employment records since 1985, with detailed information on employment type,

[2] Similar cross-country occupations data was previously constructed by (Bandiera et al., 2022), who describe the data in more detail. We construct the data in a similar way and extend their analysis. We thank the national statistical agencies of the countries listed on the website below for producing and sharing the original data: `https://international.ipums.org/international/citation_stats_offices.shtml`

[3] In most cases the national statistics office provided a sample of the microdata to IPUMS, for example drawing "a systematic sample of every 10^{th} dwelling with a random start". In other cases, the entire microdata was shared, and the sampling is done equivalently by IPUMS. See the IPUMS website for details.

[4] Several countries use this classification in their national censuses. Sometimes the occupation variable is coded in the 1968 ISCO version. In these cases, they were converted to ISCO 1988 using the stata command `iscogen` (Jann, 2019).

[5] The classification allows further differentiation into 390 'Unit Groups' at the 4-digit level, although level of granularity is not available for most countries. For further details on ISCO 88 see Hoffmann (2003).

hours, salary, and a 6-digit occupation code. It also contains demographic information on the worker, such as age, gender, and education, and information on the employing firm, such as location, firm size, and a 5-digit industry classification. The data allows us to track both firms and workers over time, so we can observe the same worker in different firms or different occupations.

A major advantage of this data is the detailed occupation classification (Classificação Brasileira de Ocupação, CBO), which contains 2,511 distinct codes. For example, language teachers are differentiated into 14 separate codes by the language they teach, economists by 7 subfields, and sports referees by the discipline they adjudicate. For each occupation, the CBO also provides a detailed list of tasks typically involved in that occupation. For example, being an economist involves a list of 83 tasks ranging from "develop data collection instruments", to "coordinate projects" to "show critical judgement". Tasks per occupation range from 5 (Forestry Extraction Worker) to 181 (Merchant Marine First Officer). Overall, there are 46,522 unique tasks classified by CBO.

Other Data Sources

We combine the above cross-country data on occupations with annual GDP per capita in constant PPP adjusted USD from the Penn World Tables Version 10.0 (Feenstra et al., 2015).

For Brazil, we also use data from the 2000 and 2010 population censuses, regional, annual GDP estimates, and input-output matrices. These are publicly available through the Instituto Brasileiro de Geografia e Estatística (IBGE).

Dataset and variable construction

For the analysis of census data (including from Brazil), we define occupational variety as the minimum number of occupations needed to jointly account for 90% of the workforce. This definition has several advantages for the cross-country analysis. Since we are using the ISCO-88 minor group (3-digit) of which there are only 116 possible values, most countries will have at least one worker of each occupation (either real or through measurement error). Our measure only counts occupations that constitute a meaningful share of the workforce. Second, this measure captures an intuitive notion of occupational variety, as it measures "how many jobs most people do". Finally, it can provide some comparability across different classification systems, provided that they have a similar level of granularity.[6] In Appendix section C, we show that our results are robust to using different percentage cut-offs and alternative measures of concentration, such as a fragmentation or Theil index.

For detailed regional analysis in Brazil, we aggregate the labour-contract level data from RAIS to build several panel datasets. The first is a yearly panel between 2003-2021 of 558 micro-regions - statistical areas that are usually considered as demarcating separate labour markets (Dix-Carneiro and Kovak, 2017). We generally refer to those as 'regions' and to this data as the regional panel. The granular, 6-digit CBO occupational classification, allows us to use the raw count of unique occupations appearing in at least one work contract in the region-year as our measure of occupational variety.[7] Similarly, we define a region as having an (active) industry if there is at least one employment contract with that industry classification. Regional GDP estimates are aggregated to this level from municipal estimates (IBGE).

We also construct panels of active industries by region-year and of active occupations by region-year, which we describe in more detail in section 5 below.

3 Stylised Facts

In this section we document and analyse the relationship between occupational variety and economic development by documenting a series of stylized facts. First, we combine the above-described data sources to document a robust relationship between occupational variety and economic development across countries, over time and across regions within Brazil. We then use the census micro data across countries to analyse how occupational variety intersects with other dimensions of structural transformation. Finally, we use employment data from Brazil, to study the relationship between occupational and industrial variety. It is a well known fact that industrial variety increases with development (Imbs and Wacziarg, 2003), and we show that this interacts in important ways with occupations.

All analyses are descriptive. Their goal is to establish a set of stylized facts and motivate the model presented in the next section.

[6] Figure A1 in the appendix illustrates this approach: For each country we rank occupations by their employment shares and then stack those shares from largest to smallest until cumulatively reaching 90%.

[7] Figure A2 in the Appendix shows choropleth maps of Brazil's micro-regions highlighting the large regional variety in occupations and the spatial correlation between occupational variety and economic development.

3.1 Occupational variety and economic development

We begin by documenting a close statistical association between economic activity and occupational variety. Figure 1 shows the relationship between occupational variety and economic development. Panel (a) plots our measure of occupational variety against GDP per capita for 44 countries for which such data is available in IPUMS. We use the latest census round for each country if multiple are available and match GDP data to the year in which the census was conducted. The fitted red line, a local polynomial smoothing estimate, shows a clear positive association. In the poorest countries most people perform less than 10 different occupations, while in the richest places the same 90% share of the workforce distributes over more than 50 occupations.

Panel (b) repeats this exercise using data from Brazil's 2010 census. The figure shows a striking regional heterogeneity within Brazil: both the number of occupation and GDP per capita vary over a range that overlaps to a large extent with the cross-country data. This remarkable heterogeneity makes Brazil an interesting case to study the mechanics of occupational variety and growth. The figure also displays the same positive association between economic development and occupational variety, as in the cross-country panel.

The final panel of Figure 1 shows the trajectory of 6 industrialising economies over the second half of the 19th century, during which time these countries started to experience large economic expansions. In all cases the growth trajectory is accompanied by a rise in occupational variety.[8]

We can use the regional panel constructed from the RAIS data to further corroborate pattern across Brazilian regions. Moving from the 2010 Population Census data of Panel b) of Figure 1 to RAIS has two implications: First, we now use the count of unique, 6-digit CBO occupations as our outcome measure, and second the sample is restricted to formal employees.

We run regressions of the following form:

$$\ln Occ_{rt} = \alpha + \beta_1 \ln GDP_{rt} + \beta_2 C_{rt} + \delta_t + \mu_r + \varepsilon_{rt} \tag{1}$$

, where C_{rt} controls for population size, δ_t denotes a time fixed effect and μ_r denotes a region fixed effect. The results, reported in panel A of Table 1, show that a 1 percent increase in regional GDP per capita is associated with a 0.6 percent increase in the number of unique occupations in the region. When considering only within-region changes over

[8] For a detailed analysis of the occupational structure in historical census data from Norway and the US see Modalsli (2017). He also finds that the trajectories in occupational composition are remarkably similar across the two countries.

	Dependent Variable: Log count of unique occupations (6-digit CBO)					
	Panel A			**Panel B**		
	(1)	(2)	(3)	(4)	(5)	(6)
Log GDP p/c	0.600***	0.435***	0.138***	0.0522***	0.146***	0.0645***
	(0.00837)	(0.00800)	(0.00913)	(0.00316)	(0.00679)	(0.00642)
Log unique industries				0.908***	0.778***	0.691***
				(0.00572)	(0.0184)	(0.0220)
region FE	NO	YES	YES	NO	YES	YES
year FE	NO	NO	YES	NO	NO	YES
Observations	8928	8928	8928	8928	8928	8928
R^2	0.537	0.979	0.984	0.958	0.990	0.991

Table 1: Regional regressions

time (holding fixed μ_r), a 1 percent increase in GDP is associated with a 0.14-0.44 increase in the number of occupations, indicating that GDP growth is associated with growth in occupational variety. The consistency of results across the two different data sources lends additional credibility to the robustness of those correlations.

3.2 Decomposing cross-country patterns

There are several well known structural transformation that typically occur as economies grow (see Gollin and Kaboski (2023) for a recent summary). To name some examples: economic activity moves out of agriculture into manufacturing and services (Herrendorf et al., 2014); the labour force becomes more educated in terms of years of formal schooling (Mankiw et al., 1992, Buera and Kaboski, 2012, Porzio et al., 2022); production moves from the household to the market, affecting especially women's labour force participation (Goldin, 1994, Ngai and Petrongolo, 2017); population moves from the country side into cities (Bryan et al., 2020); work shifts from self-employment to salaried as it becomes increasingly organised by firms (Jensen, 2022, Bandiera et al., 2022, Poschke, 2025); and industrial and product diversity expands (Imbs and Wacziarg, 2003, Hidalgo and Hausmann, 2009).

Is the rise in occupational variety a by-product of those changes or a feature of structural transformation in its own right? The census-micro data allows us to explore this question by splitting the workforce and analysing the occupational structure of different subgroups. We start in this subsection by looking at formal education, sectoral composition, gender, urbanisation, and firms. In the next subsection (3.3), we use data from Brazil for a more detailed analysis of occupational and industrial variety.

Human capital – a crucial input to economic development – is often measured as the average years of formal education in the workforce (Schoellman, 2012, Hendricks and Schoellman, 2018). We split the workforce of each country into four broad groups of educational achievement: no primary education, some primary education, some secondary education, and any tertiary or higher education. We then compute the occupational dispersion for each country within this group - e.g. how many distinct occupations account for 90% of workers without primary education. The four panels of Figure 2 show the cross-country relationship for the four education groups.

The first insight from the figure is an overall upward shift in occupational variety as we move from no primary to primary to secondary. Interestingly the fitted polynomial line is slightly lower in tertiary than secondary, indicating the large range of clerical, service and technical jobs available to workers with secondary education. But overall, workers with more formal education tend to work in more different occupations. This suggests that the overall rise in occupational variety associated with GDP is in part accounted for by a compositional shift: Countries with higher GDP tend to have a more educated workforce and more educated workers take up more different occupations.

The second interesting pattern is a positive relationship between GDP per capita and occupational variety within the no primary, primary and secondary group. Workers with no formal education choose from a much larger range of different occupations when living in a richer country. The same is true of workers with only primary education - although their share in the workforce will be much lower in richer countries, within that share, there is more occupational variety. This is interesting, as it indicates that the overall rise in occupational variety is not entirely accounted for, in a statistical sense, by a compositional shift across broad occupation groups.

Another (tangential) point is on measurement. Since workers either require some specialised skill to enter any occupation or develop such skills on the job, a larger variety of occupations likely indicates more overall ”know-how” or human capital in the economy. The positive associations in Figure 2 would then imply that years of formal education underestimate human capital, especially in richer, more diversified economies.[9]

Next, we report similar decompositions by sector, gender, urban residence and employment status. There results are shown in Figure 3. Panel a) shows that a much larger variety of occupations is found among non-agricultural workers than agricultural workers. In fact, agricultural work in most countries except the richest ones, is dominated by only a hand-full and sometimes a single occupation, while non-agricultural workers even in the poorest countries choosing between several occupations.

[9] A similar point has been made theoretically by Jones (2008).

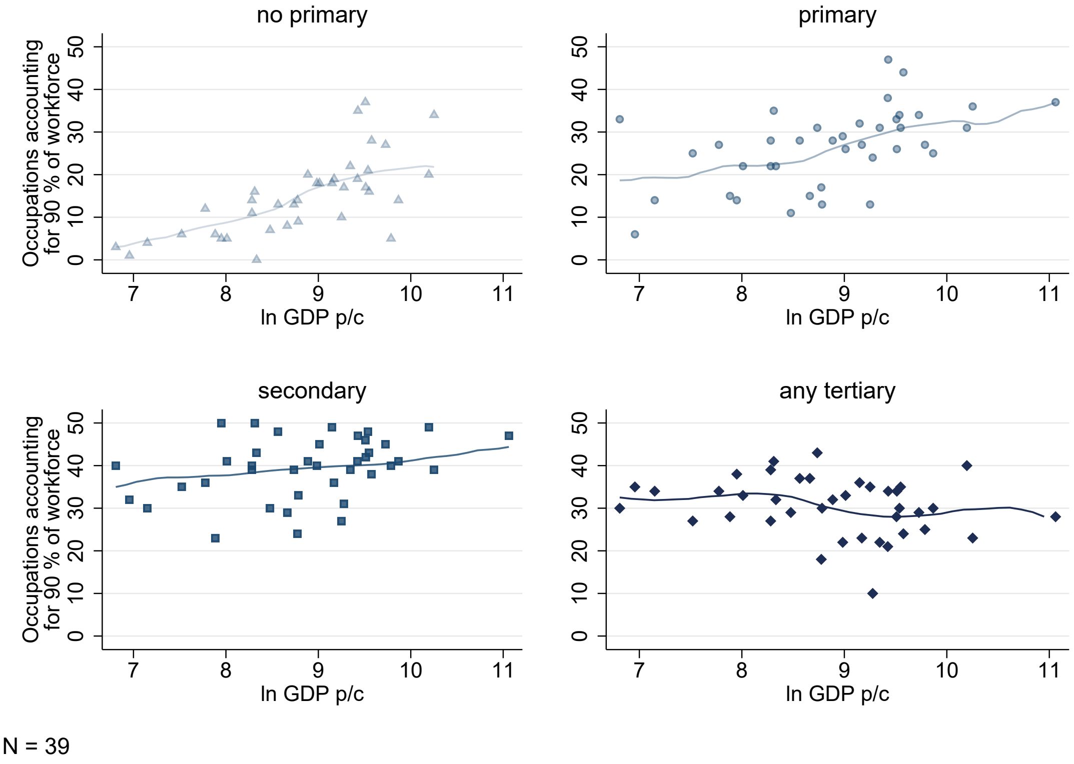
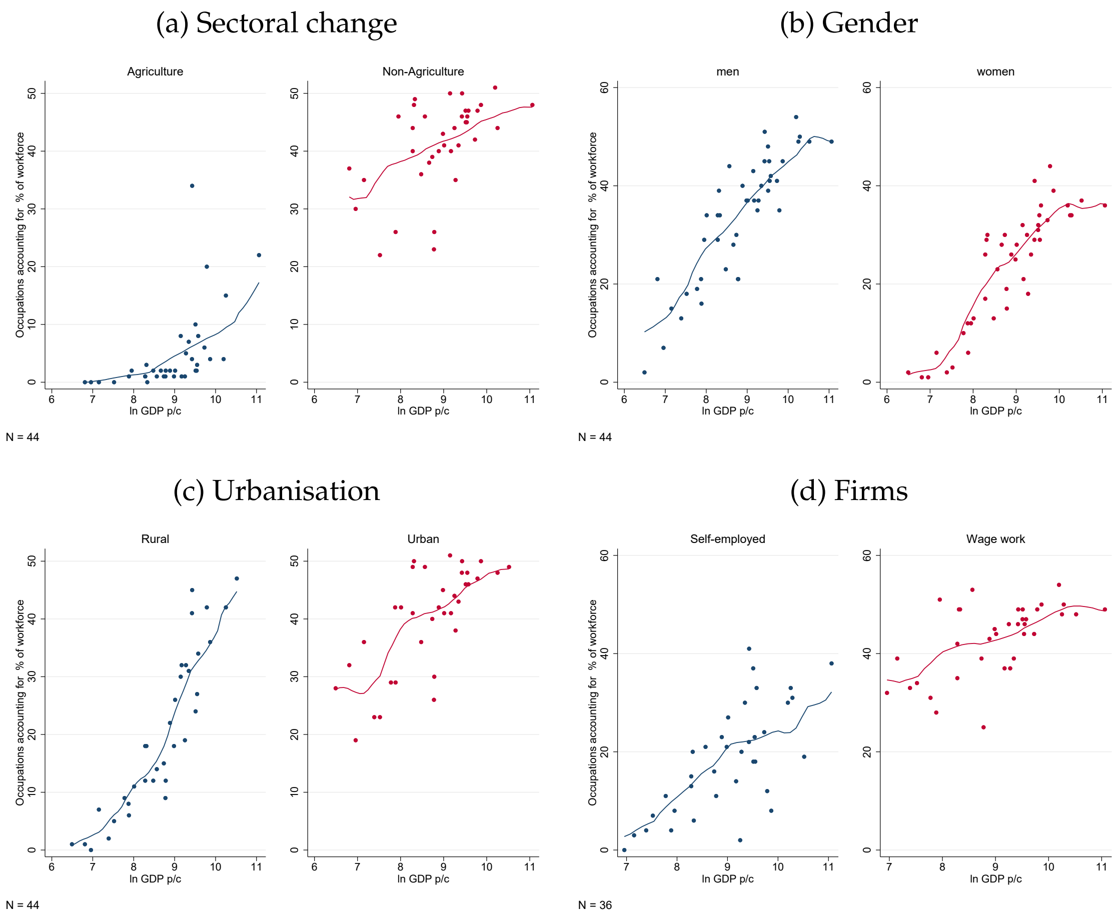


Figure 2: Occupational Variety by Education


(a) Sectoral change

(b) Gender

(c) Urbanisation

(d) Firms

Figure 3: Occupational Variety and Structural Transformation

One reason for the limited occupational dispersion within agriculture might be that the statistical classification only allows for limited variety of agricultural occupations. Indeed, out of the 116 minor occupation groups in ISCO88, only 7 fall under the major group of “Skilled Agricultural, Forestry and Fishery Workers”, while there are 22 minor groups in ”Technicians and associate professionals” and 20 in ”Plant and machine operators and assemblers”. This means that by definition, the scope for variation in agricultural occupations is limited. Part of the increase in occupations in the agricultural sector at high levels of GDP (left hand side of panel a) of Figure 3) is driven by the entry of non-agricultural occupations, such as managers, in the agriculture sector. Figure A3 in the Appendix illustrates this further. It colours the most common occupations by ISCO major group, thus highlighting that – to the extent allowed by the classification – occupational dispersion occurs both within and across broad occupation categories.

In panel b) of Figure 3, we separately report occupational variety for male and female workers. Men work in more different occupations than women at all levels of economic development, but both genders work on more occupations when living in richer countries.[10] Shifts in female labour force participation at different levels of occupational development, might explain a small part of the aggregate occupational composition. Once we include housework as a separate occupation category, it absorbs much of the female workforce and thus drastically reduces our measure of occupational variety for women (but not men). In many countries – especially at middle-income levels – it is by far the most common occupation among women, generating a U-shape for the remaining occupations that mirrors the cross-country pattern in labour force participation (see Appendix Figure A4).

Panel c) of figure 3 reports occupational variety for rural and urban workers across GDP per capita.[11] The pattern for urban workers (unsurprisingly) mirrors that of non-agricultural workers in panel (a), indicating that rural–urban migration can be a way for workers in low-income countries to access more diversified labour markets. These potential gains from rural–urban migration are much lower in richer countries: occupational variety in rural areas shows a staggering increase from the lowest level to a level that is comparable with urban areas in the richest countries.

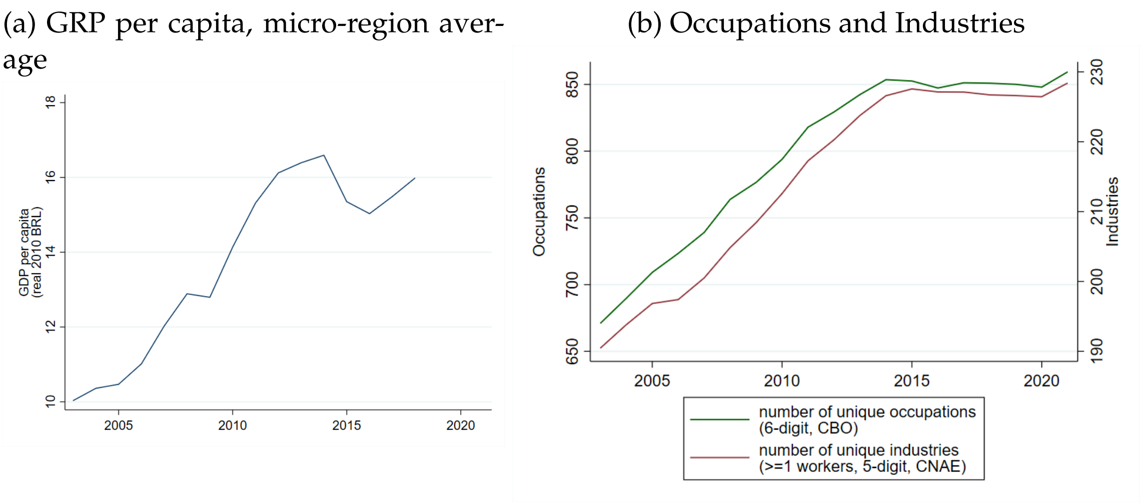
Finally, panel d) of Figure 3 splits the workforce into self-employed workers and salaried employees. The latter display substantially higher occupational variety than the self-employed throughout the range of GDP, pointing to the crucial role that firms play in the process of occupational fractionalisation. Salaried jobs even in the poorest countries display large occupational variety, but this still increases as one moves towards higher levels of GDP. The self-employed also display an increasing trend in GDP, although observations spread more noisily around the trend at high levels of GDP.

3.3 Occupations and Industries in Brazil

In this section we move to the Brazilian data to study in detail the interplay between occupations and industries. During our study period, Brazil experienced rapid economic growth, which is reflected in the rise in average gross regional product (GRP) across micro-regions (Figure 4, panel a)). With the rise in economic activity, regional economies saw an increase in both occupational and industrial variety (Figure 4, panel a)). The number of distinct industries and distinct occupations in a region evolve closely together, suggesting that the two emerge jointly as a region develops.

[10] Interestingly, one side-effect of an increase in occupational variety appears to be larger occupational segregation by gender. See Bandiera et al. (2022) for details.

[11] Urban residence is not defined consistently across countries. We follow country specific definitions. For details see: https://international.ipums.org/international-action/variables/URBAN#comparability_section


(a) GRP per capita, micro-region average

(b) Occupations and Industries

Figure 4: Average GRP per capita, occupations, and industries over time

To verify whether this is indeed the case, we pool the entire panel in Figure 5 and plot the number of unique occupations against the number of unique industries in a region-year. The figure yields several interesting insights. First, it illustrates the large regional heterogeneity with some regions having no variety and others spanning almost the entire range of possible industry and occupation classifications. For example, the 400 formal labour contracts signed in the region of Japurá in Amazonas State in 2005 are all in either of 11 occupation, and in 6 different industries. On the other hand, Belo Horizonte in Minas Gerais had 2,282 distinct occupations across 519 industries in 2014.

Second, the tight distribution of values around the positively-sloped regression line implies that there are no region-years that achieve high occupational variety without also increasing the number of active industries. There is a limit of how much specialisation (as captured by new occupations) can occur within the same industry. At some point, new occupations can only enter in the local economy through the entry of industries.[12]

This is interesting as it speaks to the nature of the specialisation of labour. Economists tend to hold at least two distinct conceptions of specialisation: i) dividing a given set of production tasks among more workers, so that each performs fewer tasks, and ii) Extending the overall set of tasks performed in the economy so that workers are more likely to work on a set of tasks at which they have a comparative advantage. The first might be attributed to Adam Smith, and the second to David Ricardo. The above suggests that, rather than the division of existing tasks into more specialised occupations, much special-

[12] This is confirmed when looking at the number of different occupations within the same industry. As an industry grows in terms of its workforce, it starts to have more different occupations, but this effect levels off quickly and few industries reach more than 200 different occupations (see Figure A5).

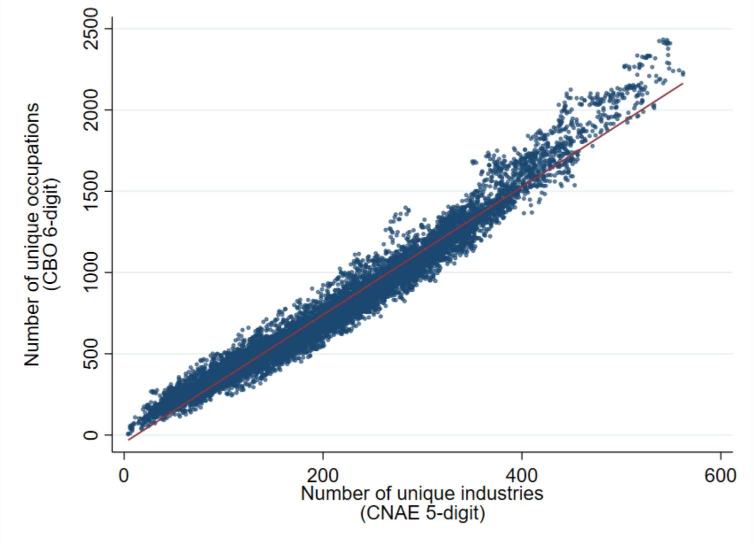


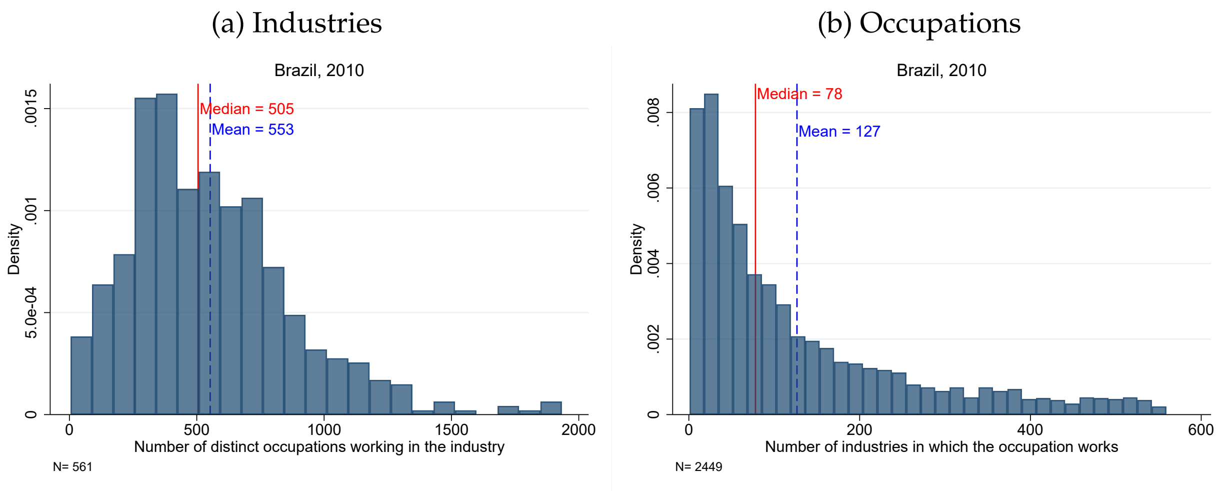
Figure 5: Correlation of occupations and industries in pooled regional sample

isation of labour is due to the introduction of new tasks with the entry of new industries (or at least the outsourcing of tasks that were previously performed in the same industry to a new one).

Finally, Figure 5 also shows that there are no region-years with large industrial variety and low occupational variety. This indicates that occupational variety indeed captures an important dimension of specialised skills required for new industry entry. It suggests that workers of different occupations are not perfectly substitutable: whenever new industries enter, they bring new occupations with them.[13]

Since new occupations emerge jointly with new industries as a region growth, the emergence of new industries explains a large part of the correlation between occupational variety and regional growth discussed in the previous subsection. Panel B of Table 1 reports the results of adding a the number of unique industries as a regressor in the region-panel regression model (equation (1)). The increase in industries accounts to a large extent for the correlation between occupations and GRP per capita, both in the pooled sample (column 4) and within regions over time (columns 5 and 6).

[13] Table B1 in the Appendix, reports results from panel regressions to confirm that the pattern described in Figure 5 is broadly robust to controlling for populations size, as well as year and region fixed effects.


(a) Industries

(b) Occupations

Figure 6: Occupational composition of industries and employment profile of occupations

From these statistics it might seem that measures of occupational and industrial dispersion capture one and the same economic phenomena. However, this is not the case, since many industries require multiple occupations while some occupations are employed across many industries. For example, when aggregating across all micro-regions in 2010, the median industry employs 505 different occupations. Figure 6 shows that there is substantial variation with some industries employing only a handful of applications and others thousands.[14] The employment of multiple occupations means that there is often overlap in the occupations that any two industries employ, creating a network where industries are linked through their use of common occupations.

On the other hand, workers of the same occupation often work in many different industries (panel b) of Figure 6), but there is large variety in how widely or narrowly an occupation is employed.[15] Their employment in different industries means that occupations overlap in the industries in which they can work, creating a network where occupations are linked through being employed in common industries.

[14] For example, in cocoa cultivation there are labour contracts with 196 distinct occupation codes, whereas we find 505 occupations in wholesale trade of meat products, 716 in telecommunications, and 885 in manufacture of organic chemicals.

[15] For example, particle physicists are seen working in only 2 industries, carpenters and sound technicians in 80 industries, chemical engineers and databank administrators in 320 industries, and production managers, accountants and passenger car drivers in almost all (>500) industries. Appendix figure A6 plots occupations' industries against their employment size, and shows that there is no strong correlation between how large an occupation is and in how many industries it works.

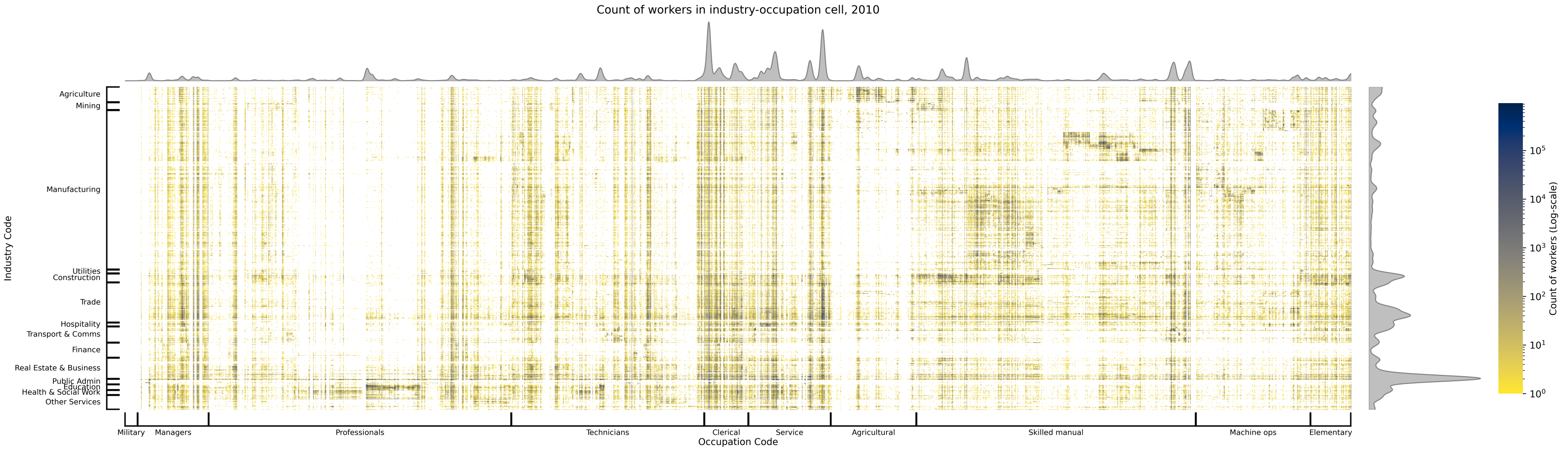


Figure 7: Occupation-Industry employment matrix, Brazil 2010

An alternative way to illustrate this network structure is Figure 7 which shows a heatmap of the distribution of workers across industries and occupations, again pooling all of Brazil in 2010. Each row refers to one of 588 industries and each column to one 2,511 occupations. Cells are shaded in a darker colour according to the number of workers in this occupation-industry combination. The colour scale is logarithmic to retain visibility of less populated cells. The gray plots at the top and right of the heatmap show marginal kernel density estimates across occupations and industries, respectively. Occupations and industries are sorted by classification and grouped into broad categories.

The figure shows a chequered pattern due to the prevalence of dark vertical and horizontal lines. The horizontal lines identify industries that hire workers of many different occupations. This is common, for example in the construction and trade sectors. The vertical lines identify occupations that are employed in many different industries. This is predominantly the case among clerical and service workers, as well as managers. Other occupations are much more clustered in a small set of industries, such as agricultural workers in agriculture, skilled manual workers in manufacturing and professionals (teachers) in education.

These descriptives suggest a network structure where industries are connected to each other through common occupational requirements, and occupations are connected to each other through being employed by common industries. In the next section, we develop a theory that formalizes this network structure, in order to understand it's implications for the entry of new industries and occupations.

4 Theory: Industrialisation with occupations

Our theoretical framework captures two main mechanisms. The first is that a larger availability of different occupations available in the local economy leads to productivity gains, as it allows workers to specialise in a set of tasks at which they have a comparative advantage.[16]

The possibility of getting a job in an occupation that matches the worker's talents, also creates incentives to obtain the specialised training needed to enter such an occupation.

[16] This mechanism lends heavily from the literature on occupational choice of heterogeneous workers and productivity gains from comparative advantage (McFadden, 1972, Eaton and Kortum, 2002, Costinot and Vogel, 2015, Costinot et al., 2016, Hsieh and Klenow, 2009). Our approach particularly builds on Hsieh et al. (2019), who present a model of occupational choice with frictions for different demographic groups (e.g. discrimination against women or black men working as lawyers). They estimate the aggregate productivity gains for the US that resulted from the decline in such frictions since the 1960s. Our theory differs in that the number of available occupations is not fixed. The main barrier of entry is not discrimination, but that in a low-variety labour market the most suitable occupation for a given worker simply doesn't exist.

Thus a larger variety of occupations encourages a larger fraction of the workforce to seek education and training (e.g. Kremer (1993), Acemoglu (1997)).

The second channel explains how new occupations can emerge as the economy grows. This part of the model draws on a literature that models industrialisation as a coordination problem (Rosenstein-Rodan, 1943, Murphy et al., 1989, Rodriguez-Clare, 1996, Rodrik, 1996, Acemoglu, 1997). In our case, the coordination problem arises as workers who train in a given occupation face uncertainty over whether they will find employment in that occupation. If there are no firms that hire this occupation in the local market, workers' expected benefits from training in it are low. On the other hand, firms won't enter a market where the specialised labour they need for production is absent or prohibitively expensive. Thus new industries only emerge if sufficient trained labour is already available.

Hence, the emergence of new occupations in our model is linked to the emergence of new industries. This aligns closely with the empirical results in the previous section showing joint emergence of new industries and occupations. However, our definition of an industry in the model is broad. The main point is that new tasks are being introduced in the economy. New occupations enter in order to perform these new tasks. This contrasts with other conceptions of "specialisation" where a given set of tasks is divided into smaller bundles.

4.1 Set-Up

There is a set of occupations $\mathcal{N}$, indexed by $n = \{1, ..., N\}$. An occupation is characterised by a fixed set of tasks, and specific skills required to perform those tasks. We assume that tasks are unique to one occupation, since empirically there is little overlap in the tasks performed by different occupations and, more importantly, we don't observe changes in the task-overlap over time.[17] This means that we can talk of occupations and tasks interchangeably. The skill requirement is summarized by an occupations-specific training costs, χ_n, which workers have to pay in order to enter the occupation.

There is a continuum of firms, $\mathcal{K}$ of mass 1, indexed by k. Each firm produces a unique final product using effective labour $\tilde{L}_{n,k}$ from a fixed, exogenous set of occupations $S(k)$. The number of occupations used by the firm is $\kappa_k = |S(k)|$.

To align the model with empirical results, we group firms into industries based on the occupations which they use for production. Let an industry $I \in \mathcal{I}$ be defined by the

[17] This is largely true in the RAIS data, which provides a detailed list of tasks for each occupation. There are a few exceptions, such as "Trabalhar em equipe" (work in a team) that is required in almost half of all occupations. See appendix Figure A7 for a matrix of task-overlap across occupations.

set of occupations required for its production, denotes $S_I \subset \mathcal{N}$. Then we can say that a firm k belongs to industry I if $S(k) = S_I$.[18] S_I thus creates a mapping between industries and occupations that determined by the exogenously given production technology. The empirical equivalent of this mapping is the heatmap of Figure 7.

There is a continuum of workers of mass 1, indexed by i, each providing one unit of labour. Each worker has a vector of occupation-specific talent for all possible occupations, $a_i = \{a_{i,n}\}_{n \in \mathcal{N}}$.

Time is discrete with $t = 1, 2, 3, ..$ Each worker lives for one period.

Industries and occupations are not active at every point in time. Let's denote by $K_t^a \subset \mathcal{K}$ the set firms that are active at time t. The remaining $K_t^d = \mathcal{K} \setminus K_t^a$ firms are "dormant". We can further define an industry to be active if a positive mass of its firms are active and denote the set of active industries by I^a. Finally, active occupations are the set of occupations hired by any active firms: $N_t^a \subset \mathcal{N}$ with $N_t^a = \bigcup_{k \in K_t^a} S(k)$. The set of active firms, industries and occupations are stock variables

4.1.1 Workers' occupational choice

We start by characterising the labour supply across occupations within one time period. It is determined by workers' occupational choice. Workers observe their talent vector and then can chose whether to train in any of the $n \in \mathcal{N}$ occupations. To do so, they have to pay a training cost that is inversely proportional to their talent. Workers can also chose not to train and work in a subsistence sector with a reservation utility of $\bar{u}$.

A worker i who expects a wage of w_n and probability of being hired if she trains in n as π_n has utility[19]

$$U_{in} = \frac{\pi_n w_n a_{in}}{\chi_n} \tag{2}$$

[18] This can be generalised to allow for within-industry specialisation, if instead we define an industry by a potential set of occupations, $S_I \subset \mathcal{N}$, that can - but don't all need to - be used in its production. The firms making up an industry, would in turn use only a subset of those occupations: $k \in I \implies S(k) \subset S_I$. Entry of firms using different occupations could then increase the set of active occupations within an industry up to the limit of potential occupations (as illustrated in Figure A5). We keep the simpler specification where all firms within an industry use the same occupations for expositional clarity. In either case the substantive mechanism is that the firms that enter perform some new task that were previously not done in the economy (or within the industry) and thus require the hiring of a new occupation.

[19] Alternatively, the log-utility might be more easily interpretable as

$$\log U_{in} = \log(\pi_n w_n) - \log\left(\frac{\chi_n}{a_{in}}\right)$$

where the first term captures the expected benefit from occupation n and the second term captures the cost of entry, which is declining in worker's talent.

The worker's hiring probability, $\pi_n \in [0,1]$, will depend on the mass of active industries that hire workers from occupation n.

Following a previous literature on comparative advantage and occupational choice (Eaton and Kortum, 2002, Hsieh et al., 2019), we assume that

$$a_{in} \overset{\text{iid}}{\sim} \text{Fréchet}(1,\theta), \quad \text{with } \theta > 1$$

Then, for a given worker, the probability of choosing occupation n is given by

$$L_n^S(w_n,\pi_n) = Pr\{n = \text{argmax } U_{in}\} = \frac{(\pi_n w_n)^\theta \chi_n^{-\theta}}{\bar{u}^\theta + \sum_{s=1}^N (\pi_s w_s)^\theta \chi_s^{-\theta}} \equiv \frac{(\pi_n w_n)^\theta \chi_n^{-\theta}}{\Phi} \tag{3}$$

The numerator captures the attraction of occupation n. This is balanced in the denominator by the attraction of all other occupations and subsistence, captured by Φ. Because there is a continuous mass of workers, the law of large numbers implies that the probability of equation 3 is also the share of workers who chose occupation n. Note that when workers don't expect to be hired ($\pi_n = 0$), they won't train in this occupation.

The average productivity of workers conditional on choosing n is given by

$$\bar{a}_n(w_n,\pi_n) = \mathbb{E}\left[a_{in} \middle| i : n = \text{argmax}_{n\in\mathcal{N}}\{U_{in}\}\right] = \gamma(\theta)\frac{\chi_n}{\pi_n w_n}\Phi^{\frac{1}{\theta}} \tag{4}$$

, where $\gamma(\theta) = \Gamma\left(1 - \frac{1}{\theta}\right)$ is a constant involving the gamma function, $\Gamma(\cdot)$.

The effective labour in occupation n is then given by

$$\tilde{L}_n^S(w_n,\pi_n) = \bar{a}_n(w_n,\pi_n)L_n^S(w_n,\pi_n) = \gamma(\theta)\frac{\left(\frac{\pi_n w_n}{\chi_n}\right)^{\theta-1}}{\Phi^{1-\frac{1}{\theta}}} \tag{5}$$

Interpreting the previous expressions, we can see that a lower training cost or higher perceived benefits through higher wages or a higher hiring probability will lead more workers to select into an occupation (equation 3). This implies a lower average productivity of those workers (equation 4). Conversely, if other occupations are perceived as relatively more attractive (high Φ), fewer workers select into n and those who nevertheless do so will have a higher productivity on average. Overall, the extensive margin effect on the worker headcount outweighs the selection effect on workers' productivity and more attractive occupations draw in more effective labour (since $\theta > 1$, equation 5).

A crucial component of the model is that the wage and hiring probability are complements from the perspective of the worker. Workers accept a lower wage if they expect to be hired with higher probability.

The remaining share of workers, $1 - \sum_{n \in N^a} L_n$ chose to remain without training and work in subsistence.

4.1.2 Firms

Each firm k uses effective labour from all $n \in S(k)$ occupations to produce a final output. The production technology is Leontief:

$$y_k = min_{n \in S_k}\{b_{n,k}\tilde{L}_{n,k}\} \tag{6}$$

with $\tilde{L}_{n,k}$ the effective labour of occupation n used in firm k and $b_{n,k}$ input shares.

To produce y_k units of output, the firm hires $\frac{y_k}{b_{n,k}}$ units of effective labour (or $\frac{y_k}{b_{n,k}\bar{a}_n}$ units of labour) from occupation n. The unit variable cost is given by

$$c_k(w) = \sum_{n \in S(k)} \frac{w_n}{b_{n,k}\bar{a}_n} \tag{7}$$

Due to the Leontief production technology, this is linear in the wages of occupations used by the firm and there are no cross-firm wage spillovers in input intensity.

Firm entry requires a fixed cost of, $F_k > 0$, that is drawn from a distribution H such that $H(0) = 0$ and H continuous on $(0, \infty)$. Each individual firm thus faces uncertainty over whether entry will be profitable enough to recover the fixed cost. Since there is a continuum of firms, however, this uncertainty will translate into a mass, λ_t^e of firms that enter in each period. Since firms only differ by the occupations they require, we can denote the mass of entering firms in an industry I by $\lambda_{I,t}^e$. Once they enter the market, firms stay active forever. We thus denote the mass of already active firms at the beginning of period t as λ_t^a and the mass of all active firms at the end of the period as $\lambda_t = \lambda_t^a + \lambda_t^e$ (and $\lambda_{I,t}^a$ and $\lambda_{I,t}$, respectively for industry I).

4.1.3 Demand and firm entry

Each firm produces one unit of output, $y(k) = y = 1$, which it can sell to the world market at a normalised price of $p(k) = p = 1$.

The operating profits of a firm k in industry I are thus given by

$$\Pi_k = 1 - c_k \tag{8}$$

And profits are $\Pi_k - F_k$. If $H(x) > 0$ for $x > 0$, a positive mass of firms enters if and only if $c_k < 1$. By the law of large numbers, the mass of new entrants is the probability of entry for any firm, ex ante, times the share of firms that are still dormant:

$$\lambda_t^e = (1 - \lambda_t^a) H(1 - c_k) \tag{9}$$

Labour demand for workers of occupation n across all industries is therefore given by

$$\tilde{L}_{n,t}^D(w) = \sum_{I:n \in S_I} \frac{\lambda_{I,t}}{b_{n,I}} \tag{10}$$

4.1.4 Hiring Probability

We assume that the hiring probability for workers of occupation n is simply the share of active firms out of those firms that could hire occupation n. That is:

$$\pi_n(\lambda) = \frac{\lambda_{n,t}^a + \lambda_{n,t}^e}{\Lambda_n} \tag{11}$$

, where $\Lambda_n = \int_{k \in \mathcal{K}} \mathbb{I}[n \in S(k)] dk$ is the mass of all firms that hire n. The denominator adds the mass of firms of all industries that hire workers of occupation n, including already active ones, $\lambda_{n,t}^a = \sum_{I:n \in S_I} \lambda_{I,t}^a$, and those expected to enter in this period, $\lambda_{n,t}^e = \sum_{I:n \in S_I} \lambda_{I,t}^e$.

4.1.5 Labour Market Clearing

Equation 5 provides the mass of workers who train in occupation n for any given wage w_n and hiring probability π_n. Since there is imperfect matching, if $\pi_n < 1$ not all of those who train will in the end get hired, indeed only the fraction π_n will. In other words, to fill one vacancy $1/\pi_n$ trained workers are needed. The labour market clearing condition is therefore

$$\tilde{L}_n^D(\lambda(w)) = \pi_n \tilde{L}_n^S(w_n, \pi_n) \tag{12}$$

This implies that there will be some over-training: some workers who trained in n won't find employment and will nevertheless work in subsistence.

Taking entry shares (and quantity) as given, we can solve for the wage that would clear the labour market:

$$\tilde{L}_n^D(\lambda_t) = \pi_{n,t} \gamma(\theta) \left(\frac{w_{n,t} \pi_{n,t}}{\chi_n} \right)^{\theta - 1} \Phi^{\frac{1}{\theta} - 1} \tag{13}$$

$$\Leftrightarrow w_n(\lambda_t) = \frac{\tilde{L}_n^D(\lambda_t)^{\frac{1}{\theta-1}}}{\pi_n(\lambda_t)^{\frac{\theta}{\theta-1}}} \chi_n \left(\frac{\Phi^{1-\frac{1}{\theta}}}{\gamma(\theta)} \right)^{\frac{1}{\theta-1}} \tag{14}$$

the first fraction on the right hand side captures two competing effects of firm entry on the wage. The numerator, $L_{n,t}^D(\lambda_t)^{\frac{1}{\theta-1}}$, captures a *labour demand channel*, by which an increase in demand for a certain occupation raises its wage. The denominator, $\pi_n(\lambda_t)^{\frac{\theta}{\theta-1}}$, captures a *hiring probability channel*: more entry makes it more likely for a trained worker to find employment and therefore makes the occupation more attractive, thus shifting out labour supply and suppressing the wage. In this specification of the model, the labour supply is governed by the Frechet parameter, θ, and since we have assumed $\theta > 1$, the elasticity of the wage with respect to π is larger than with respect to L^D. Since both labour demand and hiring probability are proportional to λ, this means, that in the model an increase in the mass of active firms hiring n (either of the same or another industry) reduces its wage, thus facilitating further entry in the next period. However, empirically it is an open question which effect dominates. Under perfect competition with no entry frictions, there is only a labour demand effect, industries with common occupations compete for workers generating negative externalities, and growth in one industry would inhibit entry of those using the same workers. We return to this discussion in the empirical section below.

4.1.6 Timing and Equilibrium

At the beginning of each period, the set of active firms, K_t^a (or equivalently, the shares of active firms in each industry $\lambda_{I,t}^a$), and associated active occupations, N_t^a, are known to everyone. The period's cohort of workers draws an idiosyncratic talent shock, a_{in} and dormant firms draw a new fixed cost, F_k.

Based on this, workers and industries make predictions about a set of wages, w_t, and hiring probabilities, π_t, that will clear the labour market for all active and entering occupations. Workers make their training decisions, generating effective labour shares, and firms make entry decisions. Since each individual firm and worker has a mass of zero, any individual entry or training decision does not affect the overall equilibrium outcome. There is hence no feedback from the entry of one individual to occupational choice and wages, and hence no strategic interaction between firms considering entry or workers considering training.

The mass of newly entering firms equilibrates to rationalise the hiring probabilities and wages.

Formally, a within-period equilibrium consists of a set of entry shares $\{\lambda_{I,t}^e\}_I$, hiring probabilities $\{\pi_{n,t}\}_n$, and wages $\{w_{n,t}\}_n$ that satisfy the entry condition (9), the hiring probability (11), and labour market clearing (12), where labour supply is determined by (5) and demand by (10).

The labour market clearing condition (12) can be solved for $\{w_{n,t}\}_n$, given any entry shares and hiring probabilities (Equation 14). Denoting the vector of entry shares across all industries by $\boldsymbol{\lambda}_t^e = \{\lambda_{I,t}^e\}_{I \in \mathcal{I}}$, we can therefore write the unit costs in (7) as a function of $\pi(\boldsymbol{\lambda}_t^e)$. Using the entry share (9), we can define the function

$$\mathcal{T}_I(\boldsymbol{\lambda}_t^e) = H\left[1 - c_I(w(\boldsymbol{\lambda}_t^e))\right] \tag{15}$$

which gives the entry share for industry I as a function of all entry shares. Then we can find an equilibrium as a fixed point of this mapping: $\hat{\boldsymbol{\lambda}}^d = \mathcal{T}(\hat{\boldsymbol{\lambda}}^d)$.

Since the wage is decreasing in λ, so is $c(\lambda)$, and hence $\mathcal{T}(\lambda)$ is weakly increasing.

From this above set-up, it is clear that in active industries new firms will tend to enter in each period. For dormant industries one possible equilibrium outcome is always to stay dormant. To see this note that $\boldsymbol{\lambda}^d = 0 \implies \pi_n = 0, \forall n \notin I^a$, which in turn implies that no workers will train in these dormant occupations. With zero effective labour, output and operating profits in dormant firms are zero, hence no dormant firm can possibly recover the fixed cost and no firm enters. The existence of this non-entry equilibrium highlights the coordination problem at the heart of our model. To rule out, indeterminacy in the case of multiple equilibria, we make an additional assumption that firms and workers are optimistic in the sense that if a positive entry equilibrium is possible, they will coordinate to reach it.

At the end of the period, newly entered firms join the next period's active firms, K_{t+1}^a, new industry enter the stock of active industries: $\lambda_{I,t+1}^a = \lambda_{I,t}^a + \lambda_{I,t}^d$ and the active mass of firms hiring any occupation n updates according to

$$\lambda_{n,t+1}^a = \lambda_{n,t}^a + \sum_{I:n \in S_I} \lambda_{I,t}^d \tag{16}$$

This means that hiring probabilities, as well as the number of active industries and occupations, weakly rise over time.

4.2 Analysis

4.2.1 More active occupations raise aggregate effective labour

The first insight from the model is that, other things equal, an increase in available occupations raises the aggregate productivity of the economy, i.e. it increases effective labour given a fixed labour endowment of 1. To see this, clearly, let's simplify the above expression for effective labour (5) by assuming symmetry across all occupation. That is, $w_n = w$ and $\chi_n = \chi, \forall n$. Let's further consider the case where $\pi_n = 1$ for all N^a active occupations and $\pi_n = 0$, otherwise. Then each active occupation attracts effective labour of

$$\tilde{L} = \gamma(\theta) \frac{\left(\frac{w}{\chi}\right)^{\theta-1}}{\left[\bar{u}^{\theta} + N\left(\frac{w}{\chi}\right)^{\theta}\right]^{1-\frac{1}{\theta}}}. \tag{17}$$

We can show that total effective labour, $N\tilde{L}(N)$ is increasing in N, by taking logs and then the derivative:

$$\frac{d\ln(N\tilde{L}(N))}{d\ln N} = \frac{d}{d\ln N}\left[\ln(N) + \ln(\gamma(\theta)) + (\theta-1) + \ln\left(\frac{w}{\chi}\right) + \left(\frac{1}{\theta}-1\right)\ln\left(\bar{u}^{\theta} + N\left(\frac{w}{\chi}\right)^{\theta}\right)\right] \tag{18}$$

$$= 1 - \left(1 - \frac{1}{\theta}\right)\frac{N\left(\frac{w}{\chi}\right)^{\theta}}{\bar{u}^{\theta} + N\left(\frac{w}{\chi}\right)^{\theta}} \tag{19}$$

which is positive for $\theta > 1$. Intuitively, increasing the number of occupations, reduces the headcount amount of labour in each individual occupation but it adds new occupations and increases the average productivity in all occupations. Having more occupations available also reduces the relative attractiveness of working in subsistence, hence increasing N^a, draws induces a larger mass of workers to obtain training training.

4.2.2 Stagnation and Entry

Next we analyse the conditions under which firms of a dormant industry can enter and thus activate the industry. As the entry condition (9) highlights, this depends on whether operating profits are large enough to offset the fixed cost. Operating profits depend (negatively) on the unit variable cost, which in turn depend (positively) on the wages of all occupations required in the industry (7) - or more precisely on the price of a unit of effective labour, $\frac{w_n}{\bar{a}_n}$, which we could call the "effective wage". Anything that affect the effective wages of its required occupations affects the chance of entry for an industry.

This highlights two insights. The first is the concept of a *bottleneck occupation*: Since there is no substitutability across occupations, a single occupation with prohibitively high wages can hold up the entry of an entire industry. An occupation is more likely to be a bottleneck if a lot of it is required in production (i.e. b_{nI} is low).

Second, more complex industries – i.e. those requiring more different occupations, or more inputs from occupations with higher training costs – are less likely to enter. As the set of occupations required by an industry, $|S_I|$, increases, it becomes more plausible that one of them creates a bottleneck.

To illustrate this, let's consider a simplified case of one dormant industry that uses only one occupation, which in turn is only used by that industry. In this case, we can derive an explicit expression of the operating profits and hence the entry share mapping (15).

These assumptions imply i) for the entry probability that $\pi_n = \frac{\lambda_{I,t}}{\Lambda}$, ii) for the unit variable cost $c_k = \frac{w_n}{\bar{a}_n b_{nI}}$, and for the effective labour demand that $\tilde{L}_n^D = \frac{\lambda_{I,t}}{b_{nI}}$.

We first substitute π_n and $\tilde{L}_n^D$ into the equilibrium wage expression (14):

$$w_n(\lambda_{I,t}) = \frac{\chi_n}{\lambda_{I,t}} \left(\frac{\Lambda^\theta \Phi^{1-\frac{1}{\theta}}}{b_{nI}\gamma(\theta)} \right)^{\frac{1}{\theta-1}} \tag{20}$$

This confirms, as discussed above, that the wage declines in the entry share.

Next, we substitute $\frac{w_n(\lambda_{I,t})}{\bar{a}_n(\lambda_{I,t})}$ in the unit cost function and collect terms to get:

$$c_k = \frac{\chi_n}{\lambda_{I,t}} \left(\frac{\Lambda}{b_{nI}\gamma(\theta)} \right)^{\frac{\theta+1}{\theta-1}} \Phi^{\frac{1}{\theta}} \tag{21}$$

Due to the direct, inverse relationship between unit cost and the entry probability, through the operating profits $(1 - c_k)$, we can see some interesting comparative statics from equation (21). In particular, since operating profits strictly increase in the share of entrants the mapping $\mathcal{T}(\lambda_I^e) = H(\Pi(\lambda_I^e))$ is strictly increasing, which means that a non-zero fixed point can exist but doesn't have to. Importantly, if $1 < c_k(\lambda_I^e)$ for all λ_I^e, even firms with the lowest fixed-cost draw wouldn't find entry profitable for any mass of entrants and the industry stays dormant. Equation (21) shows that entry becomes more likely, if the occupation has lower training cost (χ_n), the occupation has high return in this industry (b_{nI}), fewer other potential firms compete for the occupation (Λ), and fewer occupations for the workers (Φ).

4.2.3 Overlap

How, then can an initially dormant industry become active? One option are policies that lower the training cost or subsidise an industries fixed costs. However, absent such external shocks, a dormant industry can only become active if its required labour becomes more cheaply available. This in turn happens as other industries grow which hire workers from the same occupations. As the chance of being hired in an occupation rises, more workers train in those occupations, overall reducing the wage (as we saw from equation (14 above). This lower wage, in turn facilitates entry of new industries in the next period. Intuitively, if (14) is an inverse labour supply curve - that is the wage an industry has to pay to "attract" an effective labour of L_n^D - this curve becomes flatter for higher values of π_n. As other industries start to hire occupation $n \in S_I$ it becomes less likely that n in an expensive "bottleneck" occupation preventing the entry of firms in industry I.

To see this more formally, let's return to the previous example but introduce a second, active industry. Let's name the active one industry 1 and the dormant one industry 2. Let's further assume that Industry 1 has mass λ_1^a share of already active firms and that they use occupation n with the same intensity $b_{n,1} = b_{n,2} = 1$. Under these assumptions, operating profits of firms in industry 2 (the dormant industry), as a function of its own new entrants and the mass of active firms in industry 1 becomes

$$\Pi(\lambda_1^a, \lambda_2^e) = 1 - \frac{\chi n}{\lambda_1^a + \lambda_2^e} \left(\frac{\Lambda}{\gamma(\theta)} \right)^{\frac{\theta+1}{\theta-1}} \Phi^{\frac{1}{\theta}} \tag{22}$$

This shows that adding a mass of already active firms using the same occupation, reduces the wage a potential new entrant has to pay to attract workers of this occupation and thus shifts up the operating profits for any mass of potential entrants. The upward shift in $\mathcal{T}(\lambda_I^e)$ can make a fixed point with a positive mass of entrants possible, allowing industry 2 to become active.

We call this the overlap effect: If the occupational requirements of two industries overlap $S_I \cap S_{I'} \neq \emptyset$), entry of firms in one industry makes entry in the other industry more likely. This effect is stronger for industries that share a larger subset of occupations (larger $|S_I \cap S_{I'}|$) and if the share occupations are used more intensely (lower b_{nI} and $b_{nI'}$ for $n \in S_I \cap S_{I'}$).

Through the overlap effect, entry or growth of an industry in one period can trigger new entry in the next period, which in turn can trigger new entry in the period after. The overlap of occupational requirements across industries create the links in a network along which entry can thus cascade.

4.2.4 Growth dynamics

Once a positive mass of firms has entered in any industry, this facilitates further entry in future periods. This is clear from the previous discussion, as firms of the same industry act to reduce unit costs in the same way as firms from another industry that hires the same occupation. For the one-industry, one-occupation case, we can see this directly by replacing λ_I^e with $\lambda_{I,t} = \lambda_{I,t}^a + \lambda_{I,t}^e$ in equation (21). Now the firm considers entry in a period which already inherits a positive mass of active firms, λ_I^a. The operating profits, and thus the probability of entry for a given firm, increases in λ_I^a. Remember from (9) that the mass of new entrants is the remaining mass of dormant firms in the industry times the probability of entry for any dormant firm. As a larger share of firms enter in each period, the former declines while the latter increases. Initially, there are few new entrants because with low λ_t^a the probability of entry is low. As λ_t^a approaches 1 there are again few new entrants, because the share of remaining dormant firms, $1 - \lambda_t^a$ is low. Thus, once the industry becomes active and if all else remains equal, the share of active firms converges to 1 in an s-shaped way.

The economy as a whole grows through activating new industries and occupations. As the first result above indicates, more active allocations lead to a better allocation of worker talent and more training and therefore to more available effective labour. The rate of entry of new industries is determined by the occupational overlap with currently active industries. This means that growth is path dependent in this model: it occurs more easily if the current occupational structure of the economy overlaps with many new industries - or in other words, if the economy is in a denser part of the industry-overlap network.

5 Testing empirical implications

In this section, we use the RAIS micro-data to test the idea of occupation-based linkages which we have formalized above. In practice, the presence of such linkages implies that we should be able to predict which new industries enter in a region based on the region's initial occupational composition. If the required skills are already employed by currently active industries, others using those same skills will be more likely to enter. We show correlations in the panel data consistent with this narrative, and discuss some alternative explanations for these correlations. Furthermore, we explore a more plausibly causal hypothesis: growth in one industry due to an exogenous shock (e.g. foreign demand) should have a positive effect on the entry probability of industries using the same occupations. We test this hypothesis using a shift-share instrumental variable design and thus provide causal evidence of occupation based linkages.

Finally, we show that an occupational composition of the regional economy that places the region in a well-connected part of the industry-occupation network, predicts future GRP growth, even for regions at the same initial level of GRP.

5.1 Occupation-based linkages and industry entry

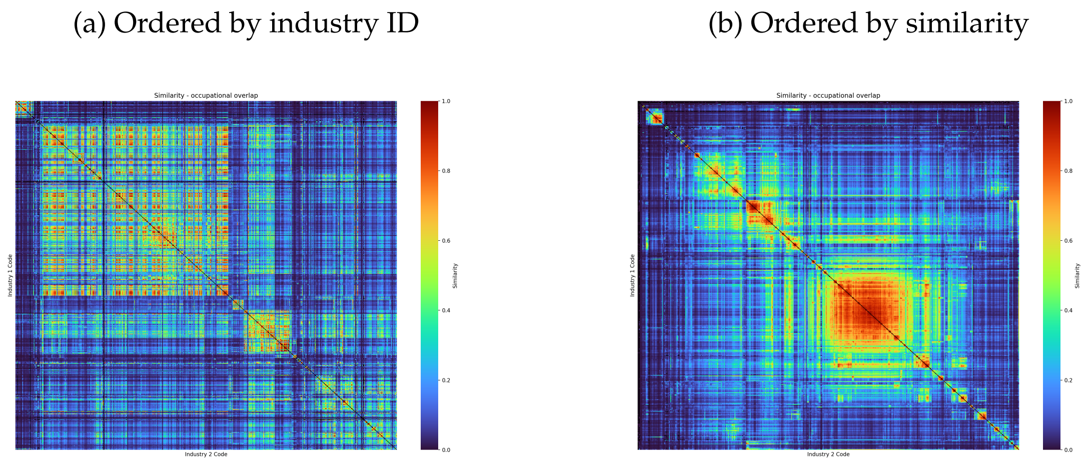
We start by defining an index of industry similarity based on industries' common use of occupations. For this we use a version of the matrix depicted in Figure 7 from a benchmark region - Sao Paulo State in 2018. We chose this because it has the most active occupations and industries in the sample - almost all of them. As in the model, we consider the similarity a fixed feature of the production technology, and hence we will hold the similarity index constant and exclude the benchmark region from the empirical analyses.

Our main measure of similarity is the cosine-similarity index, defined for any pair of industries i and j as

$$COS_{ij} = \frac{\sum_n l_{ni} l_{nj}}{\sqrt{\sum_n l_{ni}^2}\sqrt{\sum_n l_{nj}^2}} \tag{23}$$

,where l_{ni} is the count of workers of occupation n in industry i. This measure captures a relevant notion occupational overlap based on the model: two industries are considered close to each other if they have i) many occupations in common and ii) a larger share of workers in common occupations. In terms of our theoretical framework, the similarity measure is hence analogous an overlap in occupational requirements of two industries, $S_I \cap S_{I'}$ weighting for the intensity at which each occupation is used in either industry (b_{nI} and $b_{nI'}$ for all $n \in S_I \cap S_{I'}$). The cosine index falls between 0 and 1 with larger values indicating greater similarity.

Figure 8, illustrates this measure by plotting a heat map of the similarity matrix. The first matrix in panel a) leaves industries in the default order of classification. The second matrix, in panel b), orders industries so that more similar industries are clustered together, using an optimal leaf ordering algorithm (Bar-Joseph et al., 2001). This matrix displays a block structure, with several bright squares along the diagonal, indicating that the network of industries has several relatively disjoint clusters. In this, it is analogous to a the product network constructed from export data in Hidalgo et al. (2007) - which is interesting since the two networks are constructed using two entirely different concepts



(a) Ordered by industry ID

(b) Ordered by similarity

Figure 8: Industry Cosine Similarity Matrix

of similarity.[20]

With a fixed measure of industry-similarity, we now define a measure of entry-potential for every industry-region-year.

$$P_{irt} = \sum_j COS_{ij} s_{jrt} \tag{24}$$

, where s_{jrt} is the share of region r's workforce in industry j. In words, industry i has a larger potential to enter in region r and year t if this region already has many workers in many industries that are similar to i in terms of their occupational requirement. The measure hence captures an analogue to the theoretical concept of entry shares in industries using common occupations. Based on our model, we hypothesise this measure to predict industrial entry in future time periods. By construction, $P_{irt} \in [0,1]$.

To test whether this is the case we construct a panel dataset where the unit of observation is every possible industry in each region and year.[21] To test the extensive margin of entry, we define an industry as active if is at least one employment contracts (worker) in the industry in the region-year.[22]

To test the predictive power of the theory-based entry potential, we run a discrete-time

[20] In particular Hidalgo et al. (2007) use an outcomes-based approach to define two products as more similar if they are more likely to be exported by the same country. In contrast, our measure of similarity provides a micro-foundation of product similarity based on the types of specific labour typically used in production.

[21] This dataset has 6,159,762 observations, as there are 581 industry classifications, 558 micro-regions, and 19 years.

[22] We also show robustness to defining and industry as active if it employs at least 10 or 50 workers.

hazard model predicting entry of initially absent industries. This model is well suited to our context for several reasons. First, it provides an intuitive estimate of the extensive margin entry probability as a function of time and previous economic conditions in the region. Second, it correctly accounts for the censored nature of the data - many industries remain dormant in a given region throughout the panel. Finally, it allows us to consistently control for lagged explanatory variables and differential time trends across regions and industries (see e.g. Allison (1982), Singer and Willett (1993), Jenkins (1995) for details).

The hazard model is implemented by keeping only "at-risk" observations, that is industry-region cells that are not initially active and dropping all observations after an industry's first entry in a given region. We then model the probability of entry in year ($Y_{irt} = 1$), conditional on not having entered until t, with a linear probability model

$$Y_{irt} = \beta P_{ir,t-\tau} + \alpha_{rt} + \delta_{it} + \varepsilon_{irt} \tag{25}$$

The coefficient β recover a change in the entry probability (in percentage points) conditional on no prior entry, for a unit change in the regressor. A standard hazard model would include the time since the beginning of the sample t or individual year effects as regressors, to capture the baseline entry hazard. We additionally allow for region-specific baseline hazard rates, α_{rt}, and industry-specific baseline hazard rates $\delta_{i,t}$. This is possible because the entry potential varies by industry, region, and year. It is useful, as it controls for nation-wide industry effects – for example, entry of new production technologies – and region effects – for example aggregate demand spillovers of industry growth at the region level.[23]

The coefficient estimate from this model is reported in column (1) of Table 2, which indicates that similarity to existing industries in terms of occupational requirements increases the probability of entry of a new industry.[24] As our theory has highlighted, it is not obvious that occupational linkages are positively associated with the entry probability, as linked industries might compete over workers with the same skills. The fact that occupation-base linkages increase the entry probability thus lends plausibility to the mechanisms stipulated in the model by which economies grow in a path dependent manner by entering into new industries that require similar skills as existing industries.

The magnitude of the coefficient estimate is meaningful: moving from $P_{irt} = 0$ to $P_{irt} = 1$ increases the probability of entry three years later by 5 percentage points. To facilitate interpretability of this magnitude, we compute the relative increase in the entry

[23] We show that our results are robust to estimating this model with a logit link, while still including the high-dimensional fixed effects, see Appendix section B.1

[24] We use a lag of $\tau = 3$ but the results are robust for alternative lagged values, up to $\tau = 6$.

	Dependent variable: number workers in industry-region						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Industry entry potential (L3)	0.0504***	0.0398***	0.0558***	0.0534***	0.0267***	0.0306***	0.0364***
	(0.00550)	(0.00449)	(0.00423)	(0.00404)	(0.00386)	(0.00364)	(0.00372)
Direct backward linkages (L3)		0.0869***		0.103***			
		(0.0172)		(0.0167)			
Direct forward linkages (L3)			-0.176***	-0.179***			
			(0.0154)	(0.0152)			
Total backward linkages (L3)					0.0457***		0.162***
					(0.00665)		(0.0116)
Total forward linkages (L3)						0.0286***	-0.123***
						(0.00596)	(0.00868)
N	2553606	2415658	2415658	2415658	2415658	2415658	2415658
FEregXt	YES	YES	YES	YES	YES	YES	YES
FEindXt	YES	YES	YES	YES	YES	YES	YES

Table 2: Discrete-time hazard model of industry entry

probability for a 1 standard deviation increase in the regressor, P_{irt}. For the baseline specification (column (1)), we find this effect size to be 15% (se = 0.017), a large relative effect due to the small average baseline entry hazard. This estimate is robust to the inclusion of different combinations of fixed effects (Table B3), estimation with a logit link instead of the linear probability model (Table B4), and alternative definitions of an active industry as one with more than 10 or 50 workers (Tables B6 and B7).[25] The estimate is not driven by migration from surrounding regions, which itself has no effect on the entry probability (Appendix Figure A8).

5.2 Causal Identification

Does the previous estimate actually capture the causal effect of growth in one industry on the entry probability of another? We first discuss some alternative explanations and then provide direct evidence using an instrumental variable design.

5.2.1 Input-output linkages

The first alternative explanation is that input-output linkages are driving the correlation discovered above. Industry entry is facilitated by having either suppliers for intermediary

[25] We also report poisson regressions on the number of workers in each industry-region cell which show that the effect holds when including an intensive margin (appendix table B8).

inputs, or buyers of one's output already present in the local economy. Since activity of these linked industries also varies at the industry-region-year level and might be correlated with industries' occupational composition, these constitute a potential omitted variable. For example, we might observe the entry of a car manufacturing plant in a region with a growing steel industry. Our approach would wrongly attribute this to the fact that both of these industries employ mechanical engineers, while in fact entry was facilitated by the local availability of steel, an input required for car manufacturing.

To address this concern, we construct an analogous entry potential measure using industries' input-output linkages instead of occupational overlap, to compute industry similarity. Specifically, for industry i, we define the entry potential based on the share of workers in all industries that have a backward or forward linkage to industry i:

$$IO_{irt} = \sum_j L_{ji} s_{jrt} \tag{26}$$

where the strength of the linkage is taken from Brazils 2010 input-output matrix. We use four types of linkages: First, direct backward linkages from the technical coefficient matrix. These capture industries directly supplying inputs for industry i. Second, total backward linkages, from the Leontief-inverse matrix. These capture direct and indirect suppliers (i.e. suppliers of suppliers). Third, direct forward linkages. These are industries that use the output of industry i as their input. And total forward linkages, the forward equivalent of the Leontief inverse (the Ghosh inverse matrix). Columns (2) to (7) of table 2 include these measures of input-output based entry potential in the main regression and show that the occupation-based linkage remains a robust predictor after including any of the input-output linkages as controls. Backward linkages - availability of suppliers for one's inputs - consistently increase the probability of entry. Forward linkages - the availability of buyers for producers of intermediary goods - have a mixed effect. This specification also allows us to directly compare the magnitudes of the occupation-based and input-output based entry measure. As before, we gauge the effect size as the relative increase in entry probability for a one standard-deviation increase in the regressor. Taking the specification of column (5) which controls for total backward linkages, the relative effect of a 1-SD increase in P_{irt} is 8.2% (se $=$ 0.01), which is slightly smaller than the baseline estimate without controls of 15%. Compared to this, a 1-SD increase in the entry measure based on total backward linkages increases the entry probability by 13.7% (se $=$ 0.02).

5.2.2 Correlated demand shocks

A second alternative explanation is that the above results might be driven by correlated demand shocks to broad product groups. First note that the region-year fixed effect in the main specification already captures aggregate demand effects at the regional level. But this does not rule out demand shocks to specific industry groups. For example, consumers might develop a taste for meat, thus facilitating entry and growth in both the beef and pork producing industries. We might then observe growth in the beef industry, followed by entry of the pork industry, and would falsely attribute this to the fact that both industries employ butchers.

To address this concern, we define alternative measures of occupation-based entry potential that exclude occupation linkages within the same broader industry group. The idea is that we only consider occupational linkages between industries that produce different output products. In the above example, we wouldn't consider similarity between the beef and pork industry, but instead look at the effect of growth in the beef industry on industries outside the meat sector that use the same occupations. The cross-industry group entry index is defined as

$$P_{irt}^{\text{across}} = \sum_{j \neq J_i} COS_{ij} s_{jrt} \tag{27}$$

where J_i denotes industries within the same 3- or 2-digit group as i. Table 3 reports results using these alternative entry indices.[26] The first two columns replicate columns (1) and (7) of the previous table for comparison. At this 5-digit level there are 581 unique industry codes (e.g. 15.12-1: Slaughtering of poultry and other small animals and processing of meat products). In columns (3) and (4), we redefine the occupation-based entry measure by excluding linkages within the same 3-digit industry group. At this level, there are 223 unique groups (e.g. 15.1: Slaughtering and processing of meat and fish products). In the last two columns, we exclude linkages within the same 2-digit level. At this level there are 59 unique divisions (e.g. 15: Manufacture of food products and beverages).

The table shows that while the estimates become smaller as within group industries are excluded, they remain significant and sizeable, even when restricting to linkages across the 59 broad industry divisions. This suggests that effects are not mainly driven by demand shocks that are presumable correlated within these industry groups.

[26] When IO-linkages are included as control variables (in columns (2), (4), and (6)), they are defined as before without excluding any industry groups.

	Dependent variable: industry active (≥1 worker)					
	5-digit		3-digit		2-digit	
	(1)	(2)	(3)	(4)	(5)	(6)
Entry Potential (L3)	0.0504***	0.0365***	0.0399***	0.0304***	0.0153***	0.0188***
	(0.00393)	(0.00403)	(0.00398)	(0.00405)	(0.00411)	(0.00413)
Total backward linkages (L3)		0.162***		0.163***		0.162***
		(0.00816)		(0.00816)		(0.00811)
Total forward linkages (L3)		-0.123***		-0.120***		-0.115***
		(0.00792)		(0.00791)		(0.00785)
N	2553606	2415658	2553606	2415658	2553606	2415658

Table 3: Effect of linkages across broader industry groups

5.2.3 shift-share IV

Our final approach is to directly isolate exogenous shocks to one industry and check whether these affect the entry of industries to which it has occupation-based linkages.

The concern with the previous approach is that the regressor, P_{irt}, is an equilibrium outcome and might be endogenous to future industry entry (even with a lag of several years). In particular, the employment shares, s_{jrt}, might be endogenous. The industry similarity, COS_{ij}, is plausibly exogenous, as it is defined in a benchmark region at the end of the sample, which is then excluded from further analysis.

To address potential endogeneity of P_{irt}, we propose an instrumental variable strategy, using a shift-share instrument. The instrument is constructed by predicting industry j's employment share in region r as a combination of initial employment shares, $s_{jr,0}$, capturing the regional importance of industry j, and an exogenous shifter, $\Delta G_{j,t}$, capturing national expansion (or contraction) of industry j over time. The shares, $s_{jr,0}$, are held constant at their 2003 baseline level.

Instruments for P_{irt} are thus defined as

$$Z_{irt} = \sum_j COS_{ij} \times s_{jr,0} \times \Delta G_{jt} \tag{28}$$

The instrument takes exogenous changes in a linked industry j and distributes them across districts based on initial shares. Intuitively, if industry j grows at the national level, then industries i that use the same occupations should be more likely to enter in j-producing regions.

We use different industry shifts based on national aggregate employment and exports.

1. Aggregate employment: We define the first shift, $G^M_{r,jt}$, as the national employment change in industry j, leaving out j's employment in region r when computing the aggregate.

$$G^M_{r,jt} = \sum_{q \neq r} l_{jrt} \tag{29}$$

   Leaving out each region when computing it's aggregate industry shifter ensures that the aggregate is not driven by individual large regions, which would make the shifter potentially endogenous to outcomes in that region.

2. Aggregate exports: The second shifter uses aggregate exports in goods produced by industry j. We construct a regional and national exports by industry (5-digit CNAE) from municipality-level export data. This requires matching goods exports to industries using an HS4-CNAE crosswalk.[27] Again, we define a leave-one-out shifter for each region as the national total exports of all other regions:

$$G^E_{r,jt} = \sum_{q \neq r} X_{jrt} \tag{30}$$

3. Import demand: Finally, we isolate a demand shock to industry j, as the import flows of industry j's goods by other countries that do not originate in Brazil. This would, for instance, predict an increase in Brazil's meat industry if China increases its meat imports from Argentina. The intuition is that this instrument isolates growth in industry j that is driven by demand shocks in Brazil's trading partners, rather than domestic productivity shocks.

For each shifter, we define the change as the log-difference to its baseline value in 2002:

$$\Delta G_{jt} = \ln G_{jt} - \ln G_{j,2002} \tag{31}$$

This definition is useful as it cancels out unit differences. It is intuitive as it is equivalent to the cumulative log-changes since the baseline year.

Table 4 reports results of 2 stage least squares estimation using the three shift-share instruments. Panel A reports the first stage estimates and F-statistics and confirms that all three instruments are strong predictors of the occupation-based entry potential. Panel B reports the IV estimates. These range between a 2-8 percentage points, which includes

[27] Where multiple goods match to the same industry, we aggregate total exports to the industry-level. Where multiple industries match to the same export good, we distribute good exports across those industries with equal weights.

the naive OLS estimate of 5 percentage points. The most plausibly exogenous instrument based on import demand shocks, yields a marginally significant IV estimate of 2pp.

	Dependent Variable: industry active ($\geq$ 1 worker)		
	(1)	(2)	(3)
	agg. Employment	agg. Exports	Import shock
Panel A: First stage			
Instrument (L3)	1.213***	0.265***	0.000268***
	(0.132)	(0.0351)	(0.0000250)
Kleibergen–Paap F-stat	84.496	57.003	114.779
Panel B: Second stage (2SLS)			
Industry entry potential (L3)	0.0629***	0.0839***	0.0219*
	(0.00420)	(0.00785)	(0.00764)
N	2553606	2553606	2553606
FEregXt	YES	YES	YES
FEindXt	YES	YES	YES

Table 4: IV estimates of occupation-based linkages

Taken together, the results from this subsection indicate that occupation-based linkages play an important role for the entry of new industries. Entry is more likely if the occupational composition of existing industries overlaps more strongly with the occupational requirements of the potential entrant. The effect of occupation-based linkages is of a similar magnitude to that of traditional input-output linkages, it is not driven by correlated demand shocks and seems to hold in a causal identification framework using shift-share instruments.

5.3 Regional growth

In this section, we test the model's prediction that growth is path-dependent: the economy can expand into new industries more quickly if its current occupational structure places it in a denser part of the industry-occupation network. To test this, we run a simple regression in the region panel of the form

$$Y_{r,t} = \alpha + \beta_1 \bar{P}_{r,t-\tau} + \beta_2 C_{r,t-\tau} + \beta_3 Y_{r,t-\tau} + \varepsilon_{rt} \qquad (32)$$

, where $Y_{r,t}$ is gross regional product per capita, $P_{r,t-\tau}$ is the (lagged) industry entry potential index, defined in the previous section, aggregated to the region level. We use two ways of aggregation, first the average index in of all across all industries (including already ac-

	Dependent Variable: log GRP p/c					
	avg. all industries			avg. dormant industries		
	(1)	(2)	(3)	(4)	(5)	(6)
Entry potential (L5)	0.220^{***}	0.167^{***}	0.153^{***}	0.146^{***}	0.0909^{***}	0.0912^{***}
	(0.0196)	(0.0210)	(0.0204)	(0.0178)	(0.0182)	(0.0179)
Backward direct linkages (L5)		0.112^{***}			0.0974^{***}	
		(0.00914)			(0.00858)	
Forward direct linkages		-0.132^{***}			-0.154^{***}	
		(0.0129)			(0.0117)	
Backward total linkages (L5)			0.249^{***}			0.138^{***}
			(0.0203)			(0.0185)
Forward total linkages (L5)			-0.286^{***}			-0.262^{***}
			(0.0196)			(0.0180)
log GRP p/c (L5)	0.866^{***}	0.816^{***}	0.826^{***}	0.880^{***}	0.827^{***}	0.843^{***}
	(0.00402)	(0.00495)	(0.00475)	(0.00368)	(0.00477)	(0.00443)
Observations	6138	6138	6138	6138	6138	6138
R^2	0.908	0.913	0.912	0.908	0.912	0.911

Table 5: Discrete-time hazard model of occupation entry

tive ones) and second, the average entry potential among yet inactive industries. $C_{r,t-\tau}$ is a control variable which captures the (lagged and aggregated) industry entry potential based on input-output linkages. In all regressions, we control for lagged GRP per capita, effectively comparing regions at same level of development (but different occupational structure). All variables are in logs. Again, we use a lag of 5 years.

The results from this regression are shown in table 5 and confirm that regions with larger average entry potential grow more rapidly over the subsequent years.

6 Conclusion

The paper introduces and tests the idea of occupational linkages between industries. This concept can help explain a striking stylized fact: that occupational variety rises systematically with income across countries, over time, and within Brazil, and that this pattern is not fully accounted for by familiar margins of structural transformation.

Occupational linkages capture specialised skill requirements of different industries when workers from different occupations possess skills that are imperfectly substitutable. A focus on occupational structure and how it changes with development can therefore illuminate the path and speed of industrialisation. The concept of industrial linkages thus gives a micro-foundation to the observed path dependence of growth along a product-space network (Hidalgo et al., 2007), and can yield tangible policy recommendations. In our framework, traditional industrial policy and education/training policy are complements: promoting new industries without the requisite skills, or expanding training without corresponding demand, risks coordination failures. Targeting bottleneck occupations and industries with high overlap potential can unlock cascades of entry and accelerate catch-up growth. Industries can be targeted based on their specific human-capital spillovers.

Furthermore, larger variety in available occupations likely allows heterogeneous workers to specialise in an occupation that they are good at - thus increasing the scope for allocative efficiency and gains from comparative advantage. This is a crucial feature of our model, which is yet under-explored. Future research could focus on the question of how large are the potential productivity gains from increased diversification of labour.

References

Acemoglu, D. (1997). Training and innovation in an imperfect labour market. *The Review of Economic Studies 64*(3), 445–464.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological methodology 13*, 61–98.

Bandiera, O., A. Elsayed, A. Heil, and A. Smurra (2022). Economic development and the organisation of labour: Evidence from the jobs of the world project. *Journal of the European Economic Association 20*(6), 2226–2270.

Bar-Joseph, Z., D. K. Gifford, and T. S. Jaakkola (2001). Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics 17*(suppl_1), S22–S29.

Bryan, G., E. Glaeser, and N. Tsivanidis (2020). Cities in the developing world. *Annual Review of Economics 12*, 273–297.

Buera, F. J. and J. P. Kaboski (2012). The rise of the service economy. *American Economic Review 102*(6), 2540–2569.

Chaney, T. and R. Ossa (2013). Market size, division of labor, and firm productivity. *Journal of International Economics 90*(1), 177–180.

Ciccone, A. and K. Matsuyama (1996). Start-up costs and pecuniary externalities as barriers to economic development. *Journal of Development Economics 49*(1), 33–59.

Costinot, A., D. Donaldson, and C. Smith (2016). Evolving comparative advantage and the impact of climate change in agricultural markets: Evidence from 1.7 million fields around the world. *Journal of Political Economy 124*(1), 205–248.

Costinot, A. and J. Vogel (2015). Beyond ricardo: Assignment models in international trade. *Annual Review of Economics 7*(1), 31–62.

Dix-Carneiro, R. and B. K. Kovak (2017). Trade liberalization and regional dynamics. *American Economic Review 107*(10), 2908–2946.

Eaton, J. and S. Kortum (2002). Technology, geography, and trade. *Econometrica 70*(5), 1741–1779.

Feenstra, R. C., R. Inklaar, and M. P. Timmer (2015). The next generation of the penn world table. *American economic review 105*(10), 3150–82.

Goldin, C. (1994). The u-shaped female labor force function in economic development and economic history.

Gollin, D. and J. P. Kaboski (2023). New views of structural transformation: insights from recent literature.

Hausmann, R., J. Hwang, and D. Rodrik (2007). What you export matters. *Journal of economic growth 12*, 1–25.

Hausmann, R., D. Rodrik, and A. Velasco (2008). Growth diagnostics. *The Washington consensus reconsidered: Towards a new global governance*, 324–355.

Hendricks, L. and T. Schoellman (2018). Human capital and development accounting: New evidence from wage gains at migration. *The Quarterly Journal of Economics 133*(2), 665–700.

Herrendorf, B., R. Rogerson, and A. Valentinyi (2014). Growth and structural transformation. *Handbook of economic growth 2*, 855–941.

Hidalgo, C. A. and R. Hausmann (2009). The building blocks of economic complexity. *Proceedings of the national academy of sciences 106*(26), 10570–10575.

Hidalgo, C. A., B. Klinger, A.-L. Barabási, and R. Hausmann (2007). The product space conditions the development of nations. *Science 317*(5837), 482–487.

Hirschman, A. O. (1958). *The Strategy of Economic Development*. New Haven, Conn.: Yale University Press.

Hoffmann, E. (2003). International statistical comparisons of occupational and social structures. In J. H. P. Hoffmeyer-Zlotnik and C. Wolf (Eds.), *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables*, pp. 137–158. Boston, MA: Springer US.

Hsieh, C.-T., E. Hurst, C. I. Jones, and P. J. Klenow (2019). The allocation of talent and us economic growth. *Econometrica 87*(5), 1439–1474.

Hsieh, C.-T. and P. J. Klenow (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly journal of economics 124*(4), 1403–1448.

Imbs, J. and R. Wacziarg (2003). Stages of diversification. *American economic review 93*(1), 63–86.

IPUMS (2020). Integrated public use microdata series, international. `https://doi.org/10.18128/D020.V7.3`.

Jann, B. (2019, July). ISCOGEN: Stata module to translate ISCO codes. Statistical Software Components, Boston College Department of Economics.

Jenkins, S. P. (1995). Easy estimation methods for discrete-time duration models. *Oxford Bulletin of Economics & Statistics 57*(1), 129–138.

Jensen, A. (2022). Employment structure and the rise of the modern tax system. *American Economic Review 112*(1), 213–34.

Jones, B. F. (2008). The knowledge trap: human capital and development reconsidered. Technical report, National Bureau of Economic Research.

Kremer, M. (1993). The o-ring theory of economic development. *The Quarterly Journal of Economics 108*(3), 551–575.

Lane, N. (2025). Manufacturing revolutions: Industrial policy and industrialization in south korea. *The Quarterly Journal of Economics 140*(3), qjaf025.

Liu, E. (2019). Industrial policies in production networks. *The Quarterly Journal of Economics 134*(4), 1883–1948.

Mankiw, N. G., D. Romer, and D. N. Weil (1992). A contribution to the empirics of economic growth. *The quarterly journal of economics 107*(2), 407–437.

Matsuyama, K. (1991). Increasing returns, industrialization, and indeterminacy of equilibrium. *The Quarterly Journal of Economics 106*(2), 617–650.

McFadden, D. (1972). Conditional logit analysis of qualitative choice behavior. Technical report, University of California, Berkely.

Modalsli, J. (2017). Intergenerational mobility in norway, 1865–2011. *The Scandinavian Journal of Economics 119*(1), 34–71.

Murphy, K. M., A. Shleifer, and R. W. Vishny (1989). Industrialization and the big push. *Journal of political economy 97*(5), 1003–1026.

Neffke, F. and M. Henning (2013). Skill relatedness and firm diversification. *Strategic Management Journal 34*(3), 297–316.

Ngai, L. R. and B. Petrongolo (2017). Gender gaps and the rise of the service economy. *American Economic Journal: Macroeconomics 9*(4), 1–44.

Papageorgiou, T. (2022). Occupational matching and cities. *American Economic Journal: Macroeconomics 14*(3), 82–132.

Porzio, T., F. Rossi, and G. Santangelo (2022). The human side of structural transformation. *American Economic Review 112*(8), 2774–2814.

Poschke, M. (2025). Wage employment, unemployment and self-employment across countries. *Journal of Monetary Economics 149*, 103684.

Rodriguez-Clare, A. (1996). The division of labor and economic development. *Journal of Development Economics 49*(1), 3–32.

Rodrik, D. (1996). Coordination failures and government policy: A model with applications to east asia and eastern europe. *Journal of international economics 40*(1-2), 1–22.

Rosenstein-Rodan, P. N. (1943). Problems of industrialisation of eastern and south-eastern europe. *The economic journal 53*(210/211), 202–211.

Schoellman, T. (2012). Education quality and development accounting. *The Review of Economic Studies 79*(1), 388–417.

Silva, J. S. and S. Tenreyro (2006). The log of gravity. *The Review of Economics and statistics 88*(4), 641–658.

Silva, J. S. and S. Tenreyro (2011). Further simulation evidence on the performance of the poisson pseudo-maximum likelihood estimator. *Economics Letters 112*(2), 220–222.

Singer, J. D. and J. B. Willett (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of educational statistics 18*(2), 155–195.

Stammann, A. (2017). Fast and feasible estimation of generalized linear models with high-dimensional k-way fixed effects. Technical report, arXiv preprint arXiv:1707.01815.

Stammann, A., F. Heiss, and D. McFadden (2016). Estimating fixed effects logit models with large panel data. Technical report, Kiel und Hamburg: ZBW-Deutsche Zentralbibliothek für ....

Tian, L. (2021). Division of labor and productivity advantage of cities: Theory and evidence from brazil. Technical report, CEPR Discussion Paper No. DP16590.

Appendix

A Additional Figures

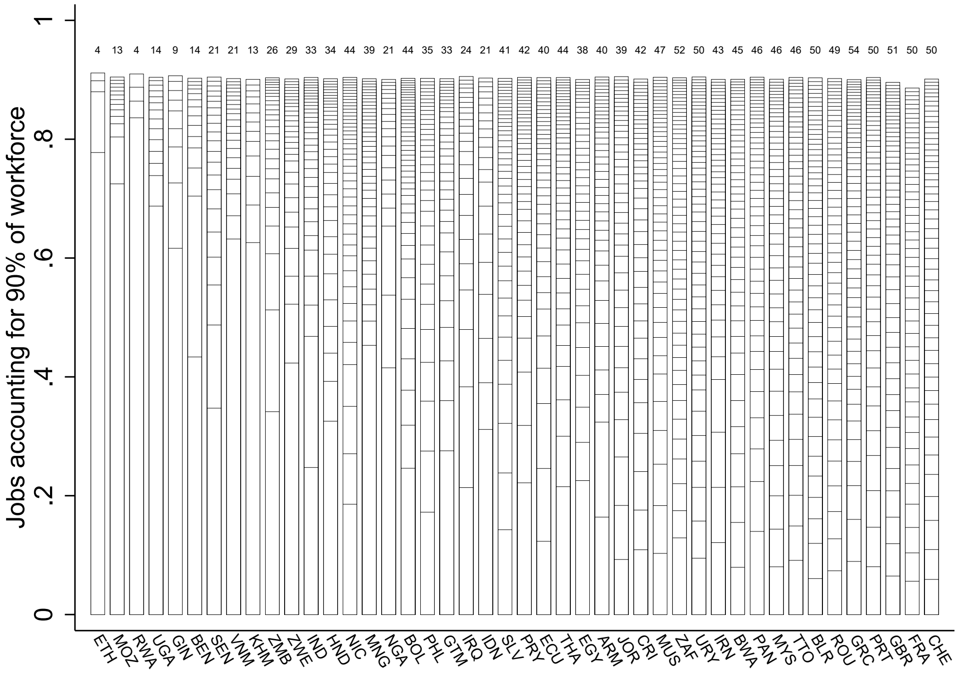
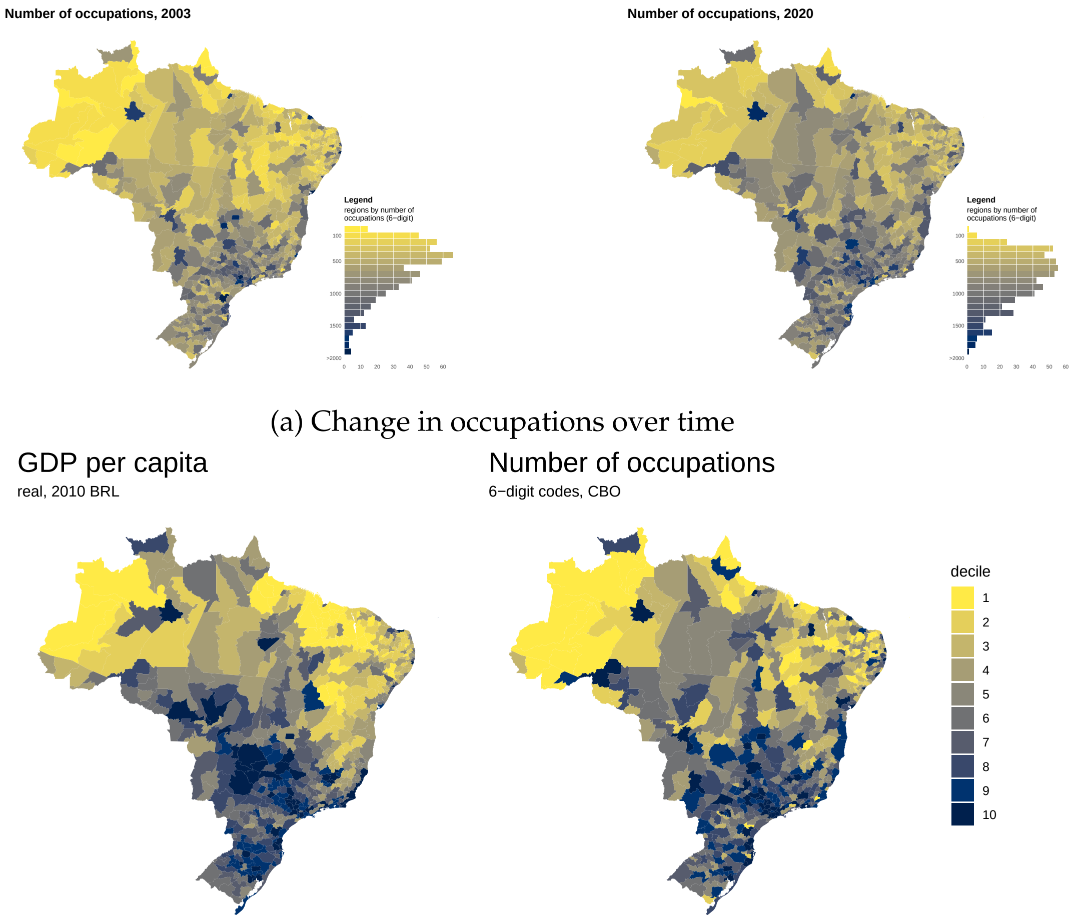


Figure A1: Stacked occupation shares up to 90% of the workforce

Notes: Every vertical bar represents an occupation defined at the ISCO88 minor group (3-digit). Bars are stacked until they reach at least 90% of the workforce of each country. Countries are sorted from left to right by increasing GDP per capita in the year in which the census was conducted.



(a) Change in occupations over time

(b) GDP per capita and occupations by decile, 2010

Figure A2: Number of unique occupations across Brazilian micro-regions

Notes:

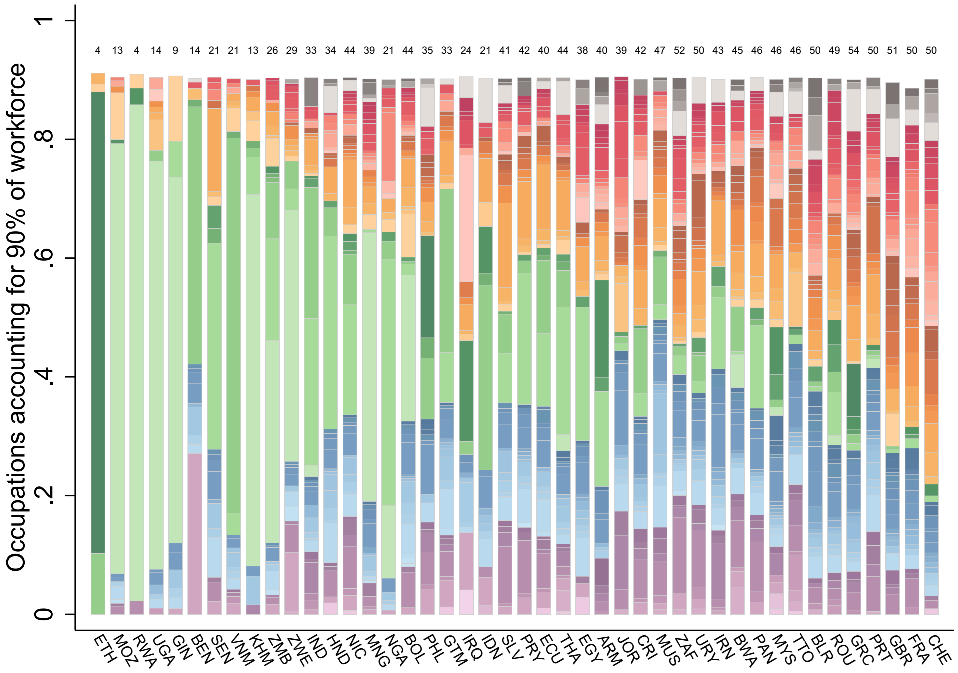


Figure A3: Stacked occupation shares up to 90% of the workforce

Notes: Every vertical bar represents an occupation defined at the ISCO88 minor group (3-digit). Bars are stacked until they reach at least 90% of the workforce of each country. Countries are sorted from left to right by increasing GDP per capita in the year in which the census was conducted. Occupations are colour-coded based and grouped by ISCO major group: administrative and legislative in grey, professional and technical in red, clerical and other services in orange, agricultural and primary in green, crafts and machine operators in blue and elementary in purple. Different shades correspond to different occupations within these groups with darker shades assigned to occupations that typically have higher educated workers (global average).

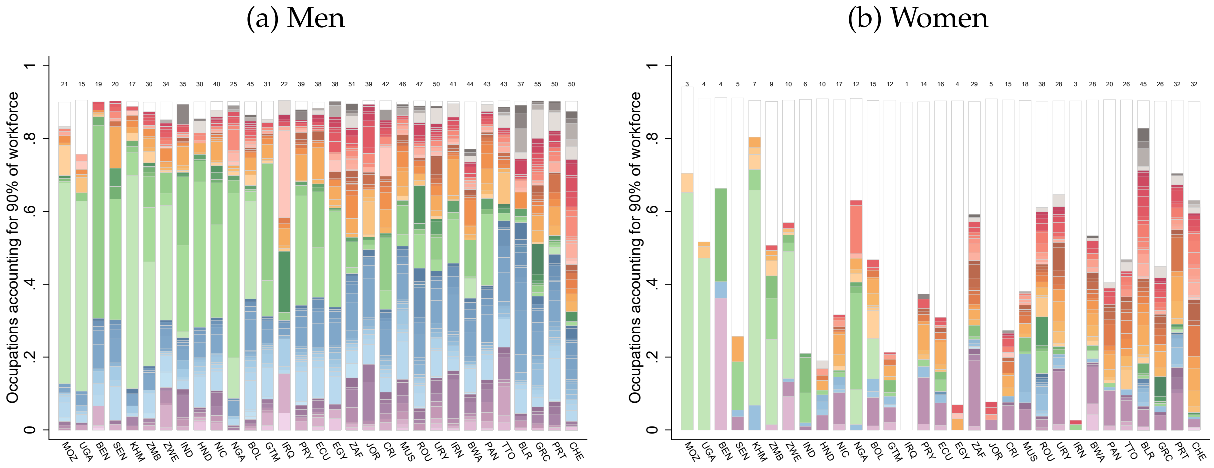


Figure A4: Occupational variety by gender

Notes:

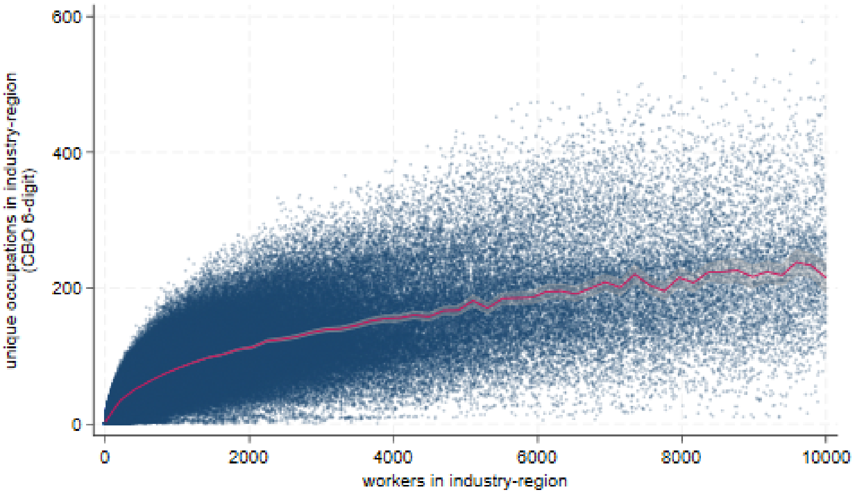


Figure A5: Occupational diversification within industries

Notes: Figure A5 (and its regression equivalent in Appendix table B1) takes as its unit of observation an industry-region-year, whenever there is a positive number of workers in this cell, and plots the number of employees against the count of unique occupations. There is substantial heterogeneity in industry size and occupational variety within industry. Regression analysis (Table B1) confirms that this is not driven by differences across industries but that the same industry can employ workers from many more occupations when operating at a larger scale (in a different region or year). At the same time, there seem to be limits to within-industry specialisation, as the number of unique occupations levels off at larger industry sizes.

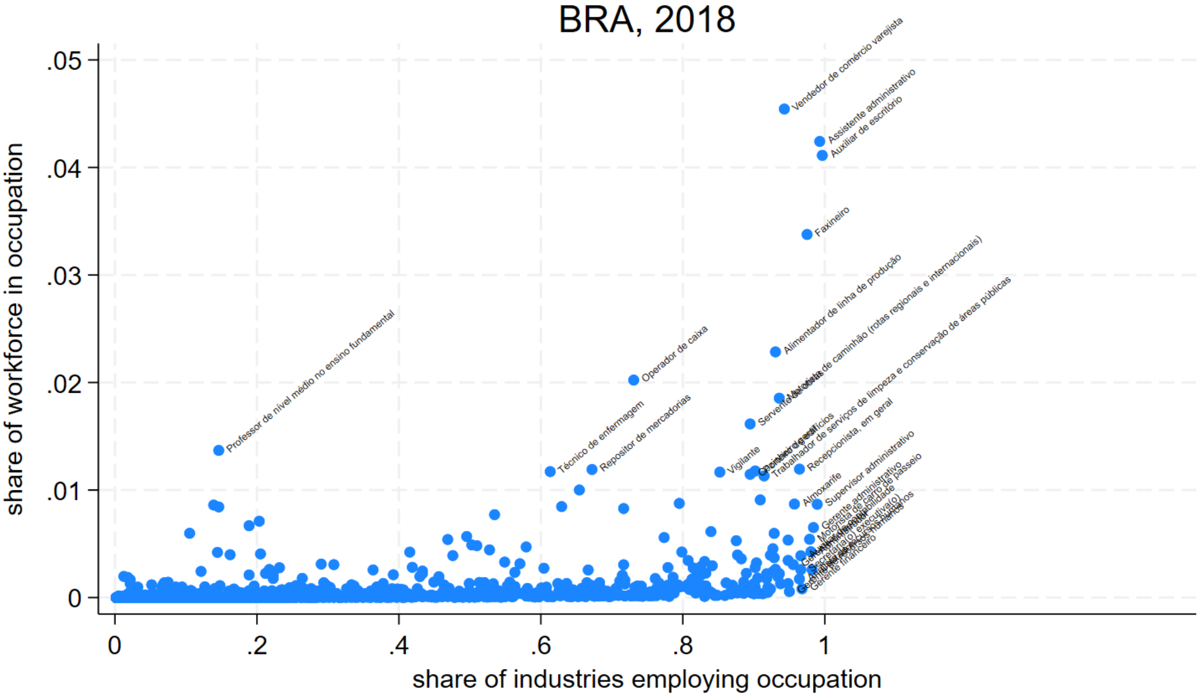


Figure A6: Occupations: size and ubiquity across industries

Notes:

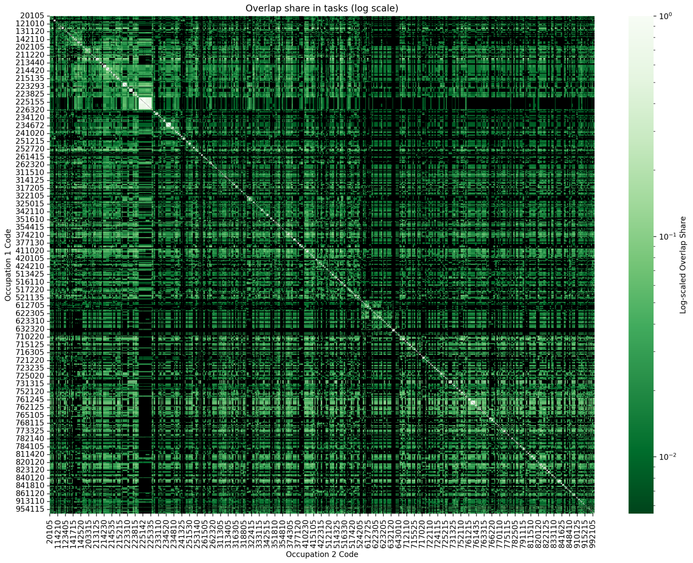


Figure A7: Heatmap of share in common tasks between occupations

Notes:

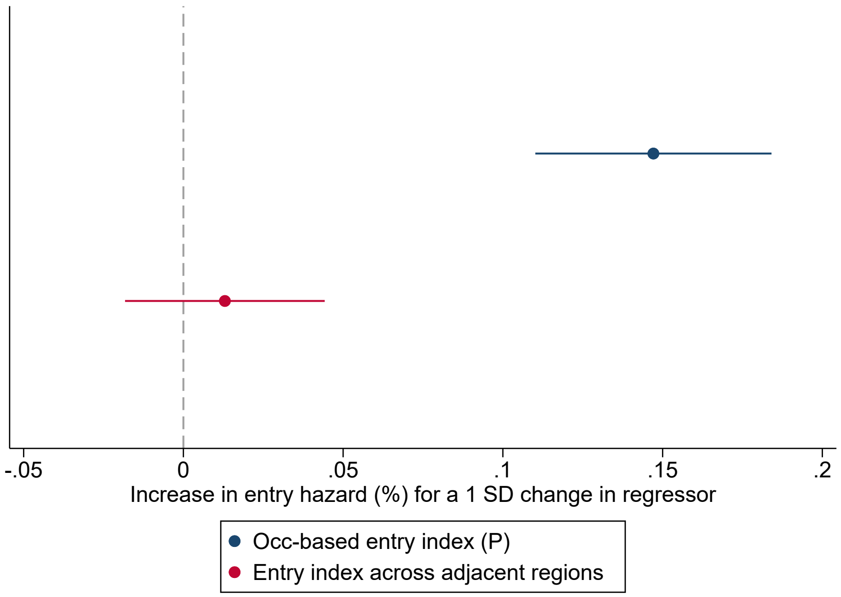


Figure A8: Effect of entry potential measure in adjacent regions (average)

Notes: Coefficients plot the relative change in entry probability for 1SD increase in regressor. We construct the occupation based entry potential measure for adjacent regions as

$$P_{irt}^{\text{adjacent}} = \sum_{j \neq i} COS_{ij} s_{j\tilde{r}t}$$

where $s_{j\tilde{r}t}$ is industry j's share of the total workforce of all regions adjacent to r.

B Additional Tables

	Dependent Variable: unique occupations			
	(1)	(2)	(3)	(4)
unique industries	3.686***	3.702***	2.545***	1.492***
	(0.0159)	(0.0159)	(0.0473)	(0.0431)
Controls	YES	YES	YES	YES
year FE	NO	YES	NO	YES
region FE	NO	NO	YES	YES
Observations	8928	8928	8928	8928
R^2	0.970	0.971	0.996	0.997

Table B1: Regions industrial and occupational variety

Notes: All regressions control for total population of the region in year t.

	Dependent Variable: log # distinct occupations (6-digit CBO)			
	(1)	(2)	(3)	(4)
Log # employees	0.617^{***}	0.600^{***}	0.605^{***}	0.603^{***}
	(0.000135)	(0.000143)	(0.000185)	(0.000187)
yearFE	NO	YES	NO	YES
regionFE	NO	YES	YES	YES
industryFE	NO	NO	YES	YES
Observations	2278147	2278147	2278143	2278143
R^2	0.899	0.903	0.934	0.934

Table B2: Within-industry occupational specialisation and industry size

Notes: The table presents results from the following regression.

$$\ln Occ_{irt} = \alpha + \beta \ln Emp_{irt} + \delta_t + \pi_r + \gamma_i + e_{irt}$$

The unit of observation is the industry-region-year (indexed by i, r, and t, respectively)

	Dependent variable: industry active ($\geq$1 worker			
	(1)	(2)	(3)	(4)
Industry entry potential (L3)	0.0412***	0.0494***	0.0398***	0.0504***
	(0.00383)	(0.00539)	(0.00331)	(0.00550)
N	2553678	2553677	2553607	2553606
FEt	YES	NO	NO	NO
FEr	YES	NO	YES	NO
FEi	YES	YES	NO	NO
FEregXt	NO	YES	NO	YES
FEindXt	NO	NO	YES	YES

Table B3: Industry entry, different FE combinations

Notes:

B.1 Logit estimation

We estimate the discrete entry hazard model discussed in section 5 using a logit model instead of the linear probability model.

$$\text{logit}(Y_{irt}) = \beta P_{ir,t-\tau} + \alpha_{rt} + \delta_{it} + \varepsilon_{irt} \tag{33}$$

In theory, the logit model is preferable as it correctly captures the binary choice nature of the industry entry problem. However, it is computationally challenging when including high dimensional fixed effects. In order to address this, we adopt the approach developed and implemented by Stammann et al. (2016), Stammann (2017). The main results are reported in tables B4 and B5. The tables report average partial effects, i.e. the percentage point change in entry probability for a unit change in the regressor. The effects range between 4-6 percentage points, consistent with but slightly larger than the effects from the linear probability model reported in the main text.

Furthermore, tables B6 and B7 use different definitions of an active industry as industries with at least 10 or 50 workers, respectively. These tables show that the result is robust to the choice of industry size.

	Dependent variable: industry active ($\geq$ 1 worker)			
	(1)	(2)	(3)	(4)
Entry potential index	0.05***	0.06***	0.05***	0.06***
	(0.00)	(0.00)	(0.00)	(0.00)
Time-at-risk (t) FE	YES	NO	NO	NO
Industry FE	YES	NO	YES	NO
Region FE	YES	YES	NO	NO
Industry×t FE	NO	YES	NO	YES
Region×t FE	NO	NO	YES	YES
Observations	2177172	2134470	1910486	1875134

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table B4: Logit model of industry entry, controlling for different fixed effects

	Dependent variable: industry active (≥ 1 worker)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Entry potential index (L5)	0.056^{***}	0.052^{***}	0.053^{***}	0.052^{***}	0.043^{***}	0.043^{***}	0.043^{**}
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003
Direct Backward Linkages (L5)		0.072^{***}		0.072^{***}			
		(0.013)		(0.013)			
Direct Forward Linkages (L5)			−0.001	−0.006			
			(0.010)	(0.011)			
Total Backward Linkages (L5)					0.028^{***}		0.039^{**}
					(0.002)		(0.007
Total Forward Linkages (L5)						0.024^{***}	−0.01
						(0.003)	(0.008
Industry × year FE	YES	YES	YES	YES	YES	YES	YES
Region × year FE	YES	YES	YES	YES	YES	YES	YES
Observations	1875134	1749770	1749770	1749770	1749770	1749770	174977

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table B5: Logit model of industry entry, controlling for input-output linkages

	Dependent variable: industry active (≥ 10 workers)			
	(1)	(2)	(3)	(4)
Entry potential index	0.06^{***}	0.06^{***}	0.06^{***}	0.06^{***}
	(0.00)	(0.00)	(0.00)	(0.00)
Time-at-risk (t) FE	YES	NO	NO	NO
Industry FE	YES	NO	YES	NO
Region FE	YES	YES	NO	NO
Industry×t FE	NO	YES	NO	YES
Region×t FE	NO	NO	YES	YES
Observations	2733588	2613845	2207105	2115091

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table B6: Discrete-time hazard model of industry entry (≥ 10 workers)

	Dependent variable: industry active ($\geq$ 50 workers)			
	(1)	(2)	(3)	(4)
Entry potential index	0.05^{***} (0.00)	0.05^{***} (0.00)	0.05^{***} (0.00)	0.05^{***} (0.00)
Time-at-risk (t) FE	YES	NO	NO	NO
Industry FE	YES	NO	YES	NO
Region FE	YES	YES	NO	NO
Industry×t FE	NO	YES	NO	YES
Region×t FE	NO	NO	YES	YES
Observations	3278186	2893740	2366320	2094955

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table B7: Discrete-time hazard model of industry entry ($\geq$ 50 workers)

B.2 Pseudo-poisson ML estimation

We estimate a poisson model to capture the extensive intensive margin effect of the occupation-based entry potential. For these analyses, our outcome is the count of workers in an industry or occupation, in a region-year. Denote this count by Y_{irt} for industry i, region r and year t. We estimate the following model using pseudo-poisson maximum likelihood (PPML) estimation

$$E[Y_{irt}] = \exp\left\{\beta P_{ir,t-\tau} + \gamma IO_{irt} + \alpha_{rt} + \delta_{it} + \varepsilon_{irt}\right\} \tag{34}$$

where $P_{ir,t}$ is the occupation based entry potential index, whose construction is discussed in the main text, $IO_{ir,t-\tau}$ denote controls for input-output linkages, and α_{rt} and δ_{it} are region-year and industry-year fixed effects, respectively. The PPML approach is well-suited to this setting with many industry-region-year cells with zero workers and multi-dimensional fixed effects (Silva and Tenreyro, 2006, 2011).

Table B8 shows a positive and significant relationship between the lagged entry potential ($\tau = 5$) and industry size.

	Dependent variable: number workers in industry-region						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Industry entry potential (L5)	10.87***	11.21***	11.40***	10.76***	8.06***	7.97***	7.99***
	(0.19)	(0.21)	(0.19)	(0.19)	(0.27)	(0.26)	(0.26)
Direct backward linkages (L5)		25.90***		21.92***			
		(1.77)		(1.64)			
Direct forward linkages (L5)			25.80***	22.85***			
			(1.03)	(1.03)			
Total backward linkages (L5)					4.49***		-0.51
					(0.26)		(1.09)
Total forward linkages (L5)						4.65***	5.15***
						(0.25)	(1.07)
N	4366350	3835134	3835134	3835134	3835134	3835134	3835134
FEregXt	YES	YES	YES	YES	YES	YES	YES
FEindXt	YES	YES	YES	YES	YES	YES	YES

Table B8: Poisson model of industry entry, controlling for I-O linkages

Notes:

B.3 Occupation entry

According to our theory, current occupational structure – through entry of new industries – should also predict entry of new occupations. To test this, we construct a panel dataset at the occupation-region level, again for every possible occupation in every region-year.

We use the industry entry potential to define an index of occupation entry potential. According to our framework, an occupation n should be more likely to emerge if (i) the industries that employ n are more likely to enter and (ii) n is more intensely used by those industries. Hence, we define

$$O_{nrt} = \sum_j P_{jrt} \bar{s}_{nj} \tag{35}$$

, where $\bar{s}_{ni} = \frac{l_{ni}}{\sum_j l_{nj}}$ is the share of occupation n workers that work in industry i - that is a measure of how important industry i is for n workers. Similar to industry similarity, we compute this in a benchmark region, Sao Paulo in 2018.

We define an occupation as active if it has at least one worker in a region-year and run an entry hazard model analogous to (33):

$$\text{logit}(Y_{nrt}) = \beta O_{nr,t-\tau} + \alpha_{rt} + \delta_{it} + \varepsilon_{irt} \tag{36}$$

The resulting average partial effects are reported in table B9. They show that current occupational structure indeed predicts which occupations enter in the future.

	Dependent variable: occupation active ($\geq$ 1 worker)			
	(1)	(2)	(3)	(4)
Occupation entry potential	0.02^{***}	0.02^{***}	0.08^{***}	0.02^{***}
	(0.00)	(0.00)	(0.00)	(0.00)
Time-at-risk (t) FE	YES	NO	NO	NO
Occupation FE	YES	NO	YES	NO
Region FE	YES	YES	NO	NO
Occupation × t FE	NO	YES	NO	YES
Region × t FE	NO	NO	YES	YES
Observations	9866189	9866189	9866189	9866189

${}^{***}p < 0.01$; ${}^{**}p < 0.05$; ${}^{*}p < 0.1$

Table B9: Discrete-time hazard model of occupation entry

C Robustness of cross-country results to alternative measures of occupational variety

We define several measures of occupational variety, all of which capture the dispersion of the workforce across the 116 ISCO minor groups. Our first measure is the minimum count of unique occupations required to jointly account for X% of the workforce. Results for $X = 90$ are reported in the main text. Figure C1 additionally reports results for $X = 80$ and $X = 95$.

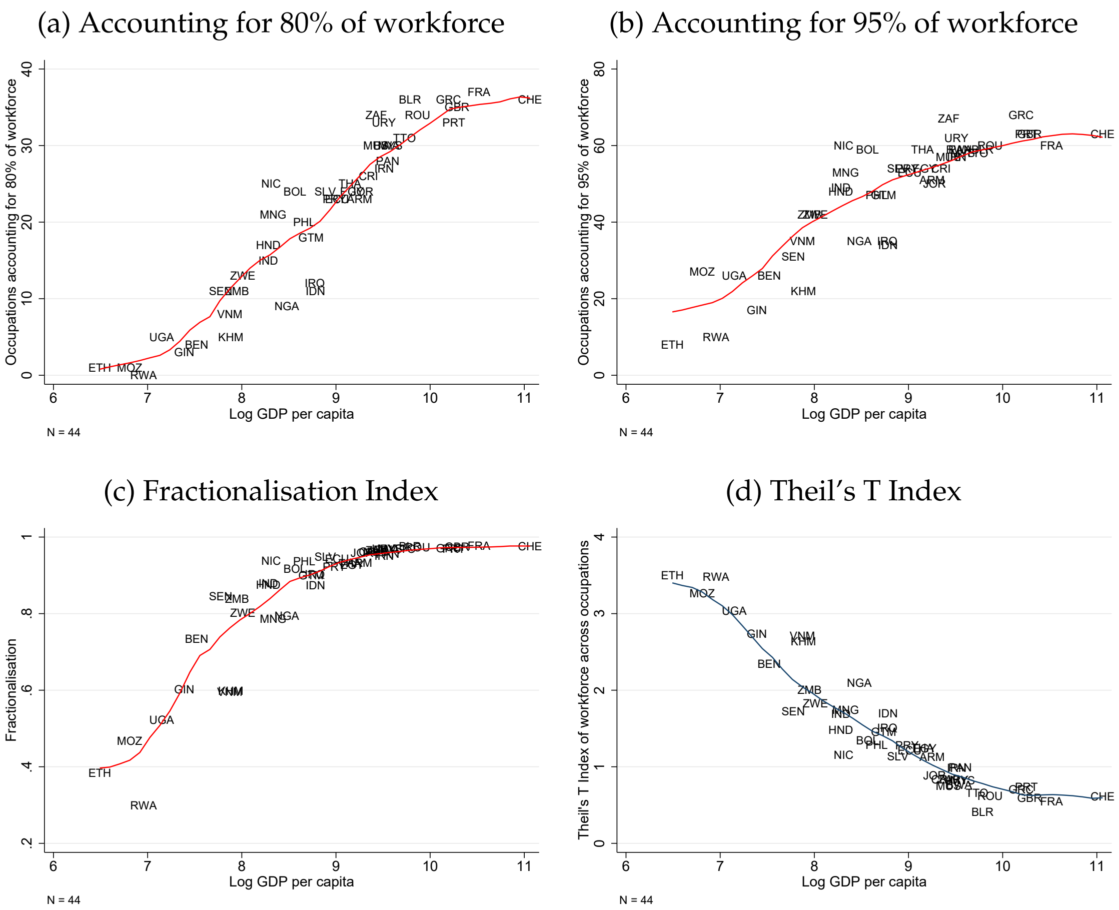
One shortcoming of this measure is that it only uses information on the largest occupation groups and discards all information outside of the X%. We consider two alternative indices that use information from all occupations. The first is a fractionalisation index, defined for country i as

$$\text{Frac}_i = 1 - HHI_i = 1 - \sum_{n=1}^{N} s_{ni}^2 \tag{37}$$

where HHI is the Herfindahl-Hirschman index across occupations, and s_{ni} is the share of country i's workforce in occupation $n = 1, ..., N$. The index ranges between 0 and 1 with higher values indicating larger fractionalisation, i.e. less concentration of the workforce across occupations. The third panel of Figure C1 plots Frac_i against Log GDP per capita. The positive pattern is robust the fractionalisation index, but it levels off as the index approaches 1 at high levels of GDP. The second index is Theil's T index, which for country i is defined as:

$$T_i = \frac{1}{N} \sum_{n=1}^{N} \frac{x_n}{\mu} \ln\left(\frac{x_n}{\mu}\right) \tag{38}$$

where x_n is the number of workers in occupation n and μ is the average number of workers across occupations. It ranges between 0 and $\ln(N)$ with larger values indicating less dispersion (more inequality) of the workforce across occupations. Accordingly, in the last panel of Figure C1, we find a clear negative association between T_i and Log GDP per capita.


(a) Accounting for 80% of workforce

(b) Accounting for 95% of workforce

(c) Fractionalisation Index

(d) Theil's T Index

Figure C1: Robustness to measures of occupational variety